Premiers pas en régression linéaire avec SAS®

Josiane Confais, Monique Le Guen

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers

Premiers pas en régression linéaire avec SAS®

Josiane CONFAIS, Monique LE GUEN

2007.47

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
http://ces.univ-paris1.fr/cesdp/CES-docs.htm
ISSN : 1955-611X

Premiers pas en régression linéaire avec SAS®

Josiane Confais (UPMC-ISUP) –
Monique Le Guen (CNRS-CES-MATISSE-UMR8174)
e-mail : confais@ccr.jussieu.fr
e-mail : monique.leguen@univ-paris1.fr

Résumé

Ce tutoriel[1] montre, de façon intuitive et sans formalisme excessif, les principales notions théoriques nécessaires à la compréhension et à l'interprétation des résultats d'analyses de régression linéaire, simple et multiple, produits par la procédure REG de SAS® et par le menu FIT de SAS/INSIGHT[2].

Ce tutoriel est issu d'un cours enseigné par les auteurs dans différentes formations : ISUP, Master de Paris 1, formation permanente du CNRS, au CEPE de l'INSEE. Il fait suite à un premier document de travail publié à l'Unité Méthodes statistiques de l'INSEE.

Nous avons ajouté de nombreux graphiques et affichages de SAS/INSIGHT qui par ses possibilités de visualisation et d'interactivité, facilite la compréhension à la fois des données et des techniques. Nous avons également ajouté des liens vers des applets ou d'autres documents accessibles sur internet.

Nous insistons dans ce tutoriel sur l'importance des graphiques exploratoires et sur les limites des résultats obtenus par une régression linéaire, si l'étape de vérification des suppositions n'est pas systématiquement entreprise.

Mots clés : Régression linéaire simple, Régression linéaire multiple, Moindres carrés ordinaires, SAS, Proc REG, SAS/INSIGHT, Graphiques exploratoires, Validation

First steps in linear regression with SAS®

Abstract

This tutorial[3] shows in an intuitive way and without excessive formalism, the theoretical notions necessary to understand and interpret simple and multiple regression produced by SAS® PROC REG and by the menu FIT of SAS/INSIGHT[4].

This tutorial is based heavily on training courses given by the authors in high-profile institutions like ISUP, Master Degree at Paris 1 University, CNRS and CEPE-INSEE. It follows a first working paper published by UMS-INSEE. Thanks to SAS/INSIGHT interactivity and visualization tools, we created numerous graphs and displays to improve the understanding of data and statistical methods. This tutorial also includes various links towards applets or other documents from the internet.

We insist in this tutorial on the significance of exploratory graphs and on the limits of results obtained by a linear regression if the assumptions are not systematically checked.

Keywords: Simple Linear Regression, Multiple Linear Regression, Ordinary Least Squares, SAS, Proc REG, SAS/INSIGHT, Exploratory Graphics, Validation.

JEL Classification **:** C01, C52 AMS Classification **:** 62J05

---

[1] Ce tutoriel est publié dans le n°33 de la revue électronique MODULAD à l'adresse http://www-rocq.inria.fr/axis/modulad/numero-35/Tutoriel-confais-35/confais-35.pdf

[2] SAS® et SAS/INSIGHT sont les marques déposées de SAS Institute Inc., Cary, NC, USA

[3] This tutorial was published in the electronics review MODULAD, n° 33, http://www-rocq.inria.fr/axis/modulad/numero-35/Tutoriel-confais-35/confais-35.pdf

[4] SAS® and SAS/INSIGHT are registered trademarks of SAS Institute Inc., Cary, NC, USA

Sommaire

1. Sensibilisation à la régression linéaire simple

Cette sensibilisation à la régression présente de manière détaillée la logique et les calculs permettant la compréhension de la régression simple.
On montre tout d'abord la démarche algébrique qui conduit à un ajustement affine, puis par un détour obligé à l'Histoire, on « glisse » vers la modélisation en s'appuyant sur la Statistique.

1.1. Où se place la régression linéaire ?

La **régression linéaire** se classe parmi les méthodes d'analyses multivariées qui traitent des données quantitatives.

C'est une méthode d'investigation sur données d'observations, ou d'expérimentations, où l'objectif principal est de rechercher une liaison linéaire entre une variable Y quantitative et une ou plusieurs variables X également quantitatives.

C'est la méthode la plus utilisée pour deux raisons majeures :

- c'est une **méthode ancienne,**
- c'est l'**outil de base** de la plupart des modélisations plus sophistiquées comme la régression logistique, le modèle linéaire généralisé, les méthodes de traitement des séries temporelles, et surtout des modèles économétriques, etc.

A l'aide du tableau 1.1, on peut repérer les méthodes les plus courantes d'analyses statistiques et les procédures SAS utiles pour rechercher des liaisons, selon le type (nominal, ordinal, intervalle, ratio) des variables Y et X. Le lecteur peu familiarisé avec la terminologie des variables SAS pourra voir sur le site de MODULAD, le tutoriel[5] « *La Proc FREQ de SAS, Tests d'indépendance et d'association* », de J. CONFAIS, Y. GRELET, M. LE GUEN.

[5] http://www-rocq.inria.fr/axis/modulad/archives/numero-33/tutorial-confais-33/confais-33-tutorial.pdf , page 5-7.

Tableau 1.1 Procédures SAS adaptées selon le type des variables (nominal, ordinal, intervalle, ratio)

	X intervalle/ratio	X ordinale/nominale	
Y intervalle/ratio	Régression linéaire ***PROC REG***	Analyse de la variance ***PROC ANOVA***	Modèles linéaires généralisés ⇦ ***PROC GLM***
Y ordinale/nominale	*Si Y est ordinale ou à 2 modalités* Régression logistique ***PROC LOGISTIC***	Analyses de tableaux de contingence ***PROC FREQ*** Régression logistique ***PROC LOGISTIC***	Traitements des variables catégorielles ⇦ ***PROC CATMOD***

Pour la régression linéaire la procédure REG est la plus complète. Cependant le module SAS/INSIGHT, qui est à la fois un tableur, un grapheur et un analyseur, est particulièrement adapté pour étudier des données dans une problématique de régression linéaire couplée à une analyse exploratoire des données.

Dans les exemples nous utiliserons l'une ou l'autre de ces possibilités. En annexe 2, on trouvera un mode d'emploi très succinct de SAS/INSIGHT.

1.2. Ajustement affine ou Régression Simple

Exemple
Soient les 2 mesures de poids (variable X) et taille (variable Y) relevées sur un échantillon de 20 objets.

Tableau 1.2 Données Taille et Poids

identifiant	poids (X)	taille (Y)
1	46	152
2	78	158
3	85	160
4	85	162
5	85	158
6	85	159
7	95	165
8	95	165
9	100	166
10	100	159
11	100	166
12	103	168
13	105	163
14	105	164
15	115	168
16	115	166
17	115	162
18	130	165
19	135	167
20	150	172

Le graphique du nuage de points, d'abscisse le **poids** et d'ordonnée la **taille** montre qu'il existe une **relation linéaire** entre ces deux variables. Lorsque le poids augmente, la taille a tendance à croître également.

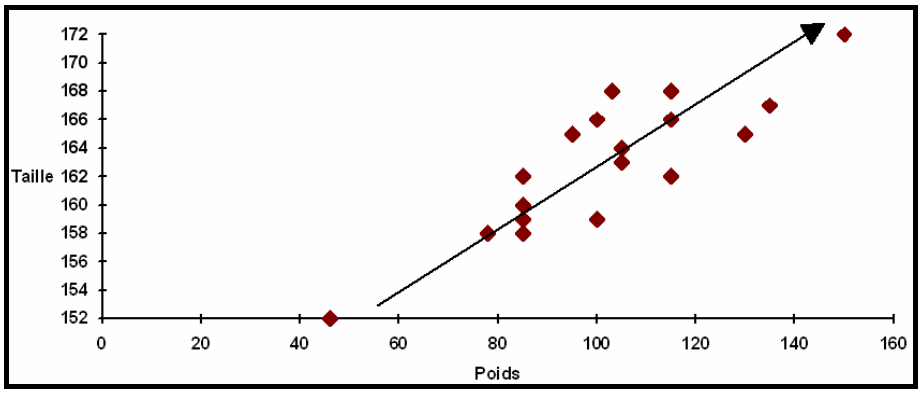


Figure 1.1 Taille*Poids

Les points du nuage sont approximativement alignés sur une droite (y=ax+b) à une erreur près.

$$\textbf{Taille} = \beta_0 + \beta_1 \textbf{ Poids + erreur}$$

La variable Taille (Y) est appelée la variable "**réponse**", ou selon les domaines disciplinaires, variable à expliquer, ou encore variable dépendante. La variable Poids (X) est la variable "**régresseur**", encore appelée variable explicative, ou variable indépendante.

β_0 est l'ordonnée à l'origine.

β_1 est la pente de la droite d'ajustement.

Note : Dans ce document nous n'utiliserons que les termes « réponse » et « régresseurs »**,** pour éviter toutes confusions sémantiques très dommageables lors des interprétations des résultats, et particulièrement lors de la communication des résultats à un tiers.

Par exemple, la variable dite expliquée n'est pas forcément expliquée par les variables dénommées explicatives. Quand aux variables dites indépendantes, elles sont, dans le cas de données réelles, rarement indépendantes.

1.2.1. Comment trouver la droite qui passe « au plus près » de tous les points?

Pour trouver la droite qui passe « au plus près » de tous les points il faut se donner un **critère d'ajustement**.

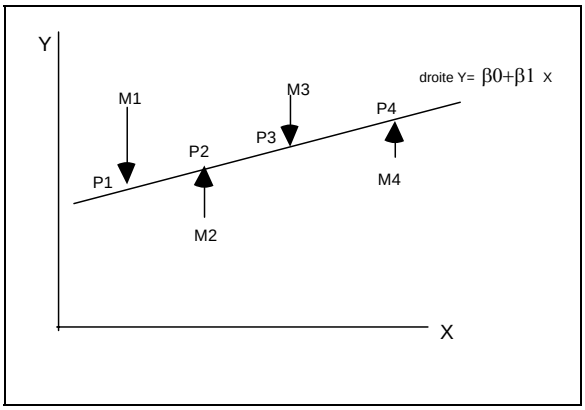


Figure 1.2 Projection des points M_1...M_4 sur la droite.

On projette les points M_1 à M_4 parallèlement à l'axe des Y. Sur la droite on obtient les points P_1 à P_4, comme le montre la figure 1.2. Le critère retenu pour déterminer la droite D passant au plus près de tous les points sera tel que :

La somme des carrés des écarts (SCE) des points observés M_i à la droite solution soit minimum.

La droite solution sera appelée **droite de régression de Y sur X**.

Le critère est le **« critère des Moindres Carrés Ordinaires »** (MCO, *Ordinary Least Squares* en anglais), appelé aussi par les statisticiens « **critère de Norme L^2** ».
Les écarts sont calculés en projetant les points M parallèlement à l'axe des Y. On pourrait aussi projeter les points M parallèlement à l'axe des X, on aurait alors une autre droite solution (régression de X sur Y). Dans ces deux régressions Y et X ne jouent pas le même rôle.

On pourrait aussi projeter les points M perpendiculairement à la droite solution. Y et X joueraient dans ce cas le même rôle. C'est la situation que l'on rencontre dans une Analyse en Composantes Principales[6], illustrée dans la figure 1.3.

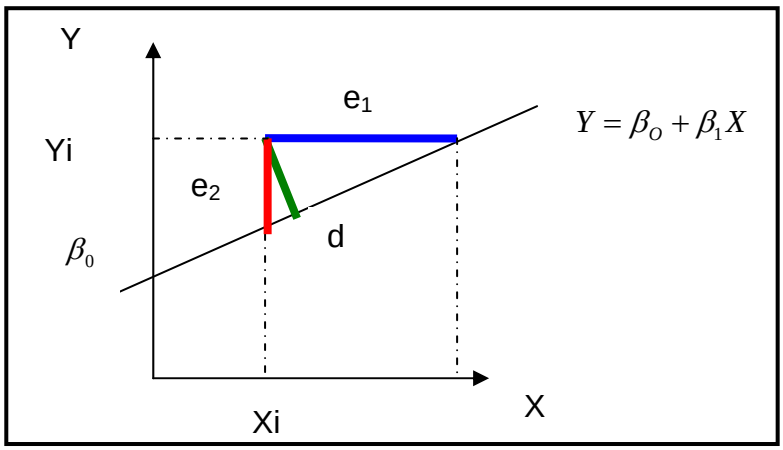


Figure 1.3 Trois projections possibles du point (X_i, Y_i)

1.2.2. Méthode d'estimation des paramètres β_0 et β_1

La Somme des Carrés des Ecarts (SCE) est donnée par :

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

La valeur de cette fonction S est minimum lorsque les dérivées de S par rapport à β_O et β_1 s'annulent. La solution est obtenue en résolvant le système :

$$\frac{\partial S}{\partial \beta_0} = 0 \quad \text{et} \quad \frac{\partial S}{\partial \beta_1} = 0$$

Les dérivées par rapport à β_0 et β_1 sont :

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_O - \beta_1 X_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_O - \beta_1 X_i)$$

Ces dérivées s'annulent pour deux valeurs b_0 et b_1 solutions des 2 équations à 2 inconnues :

[6] On pourrait encore prendre comme critère la somme des valeurs absolues des écarts des points observés à la droite, ce serait alors un critère de norme L^1, et pourquoi pas prendre un exposant non entier appartenant à l'intervalle [1,2], ce serait une norme L^p.

équation 1 : $$\sum_{i=1}^{n}(Y_i - b_O - b_1 X_i) = 0$$

équation 2 : $$\sum_{i=1}^{n} X_i(Y_i - b_O - b_1 X_i) = 0$$

Ce système de 2 équations à 2 inconnues déterminent les **équations normales.**

Développons ces 2 équations normales :

• l'équation 1 donne : $\sum Y_i - nb_0 - b_1 \sum X_i = 0$ et en divisant par n $\overline{Y} = b_0 + b_1 \overline{X}$.

On remarque que la droite solution passe par le centre de gravité du nuage $(\overline{X}, \overline{Y}) = \left(\frac{\sum_i X_i}{n}, \frac{\sum_i Y_i}{n} \right)$.

• L'équation 2 donne

$$\sum Y_i X_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

dans laquelle on remplace b_0

$$\sum Y_i X_i - (\overline{Y} - b_1 \overline{X}) \sum X_i - b_1 \sum X_i^2 = 0$$

Solution :

$$b_1 = \frac{\sum X_i Y_i - (\sum X_i \sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2 / n}$$

en divisant numérateur et dénominateur par n on retrouve les expressions de la covariance et de la variance empiriques :

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{Cov(X,Y)}{Var(X)}$$ ***formule n° 1***

Les points qui sont sur la droite de régression ont pour ordonnée: $\hat{Y} = b_0 + b_1 X$

Le coefficient b_1 dépend au numérateur de la covariance entre X et Y, et de la variance de X pour le dénominateur.

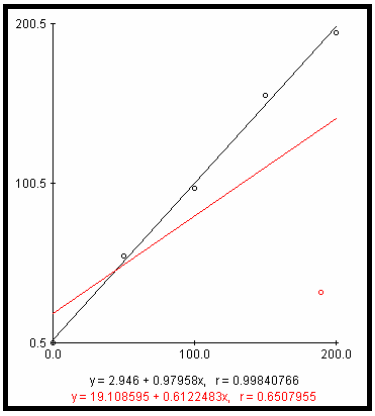
<u>Terminologie</u>

$\hat{Y}$ est l'**estimation** de Y obtenue à partir de l'équation de régression.
$\hat{Y}$ *se prononce Y chapeau.*

b_0 et b_1 sont les **estimateurs** des moindres carrés des paramètres inconnus β_0 et β_1. On appelle estimations les valeurs particulières (solutions) prises par les estimateurs b_0 et b_1.
Dans la suite du document on ne fera pas de différence de notations entre les estimateurs b_0 ou b_1 et leurs estimations.

1.2.3. Effet d'un point observation sur la droite de régression

Avec cet applet java http://www.stat.sc.edu/~west/javahtml/Regression.html
on peut voir l'effet de levier (leverage) sur le calcul de la droite de régression en ajoutant un point -rouge- par un simple clic de souris. Ici le point rouge est un point influent dans la liaison (X,Y). Plus le point est éloigné de la tendance plus son levier sera grand. Il peut aussi exister des points atypiques -Outliers- seulement en direction des X, ou dans la direction des Y (voir le chapitre 4).



1.2.4. Décomposition de l'écart entre Y_i et la moyenne de Y

En un point d'observation (X_i, Y_i) on décompose l'écart entre Y_i et la moyenne des Y en ajoutant puis retranchant $\hat{Y}_i$ la valeur estimée de Y par la droite de régression. Cette procédure fait apparaître une somme de deux écarts :

$$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})$$
$$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$

Ainsi l'écart total $(Y_i - \overline{Y})$ peut être vu comme la somme de deux écarts :

- un écart entre Y_i observé et $\hat{Y}_i$ la valeur estimée par le modèle
- un écart entre $\hat{Y}_i$ la valeur estimée par le modèle et la moyenne $\overline{Y}$.

Le graphique suivant montre l'explication géométrique de cette décomposition.
Cet artifice de décomposition aura un intérêt fondamental dans l'analyse de la variance abordée au paragraphe suivant.

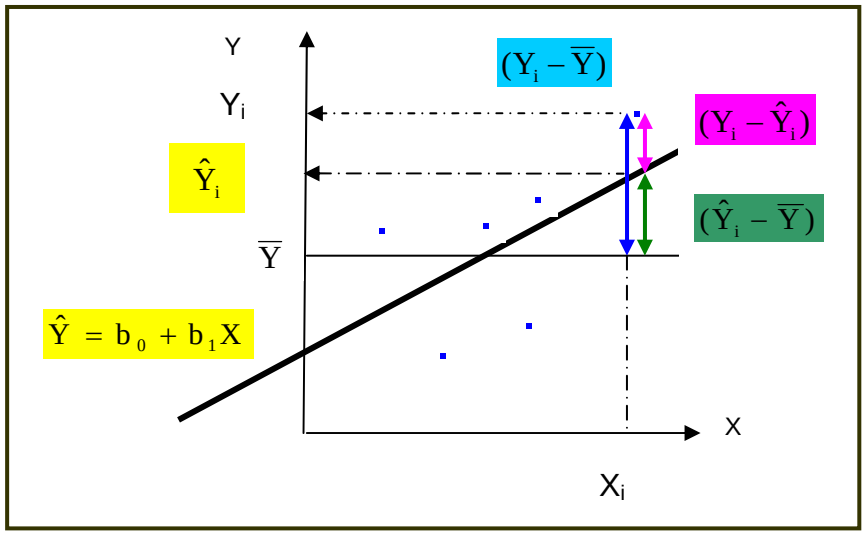


Figure 1.4 Décomposition des différents écarts

Ecart total $(Y_i - \overline{Y})$ = écart dû au model $(Y_i - \hat{Y}_i)$ + écart résiduel $(\hat{Y}_i - \overline{Y})$

1.2.5. Analyse de la variance

Ce que le modèle explique et ce qu'il n'explique pas

A partir de l'équation de la droite de régression (modèle retenu), on peut pour tout point i d'abscisse X_i calculer son estimation (ordonnée) $\hat{Y}_i$

$$\hat{Y}_i = b_0 + b_1 X_i \quad \text{avec} \quad b_0 = \overline{Y} - b_1 \overline{X}$$

ce qui donne :

$$\hat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X})$$

ou encore

$$\hat{Y}_i - \overline{Y} = b_1 (X_i - \overline{X}) \qquad \textbf{\textit{formule n° 2}}$$

En un point i l'écart ou résidu est : $Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - (\hat{Y}_i - \overline{Y})$

On élève les deux membres au carré et on somme sur les observations i :

$$\sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \overline{Y})^2 + \sum_i (\hat{Y}_i - \overline{Y})^2 - 2\sum_i (Y_i - \overline{Y})(\hat{Y}_i - \overline{Y})$$

En utilisant la *formule n°2* :

$$\sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \overline{Y})^2 + \sum_i (\hat{Y}_i - \overline{Y})^2 - 2\sum_i (Y_i - \overline{Y}) \cdot b_1 (X_i - \overline{X})$$

En utilisant une transformation de la formule n°1 : $b_1 \sum (X_i - \overline{X})^2 = \sum (X_i - \overline{X})(Y_i - \overline{Y})$ on obtient

$$\sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \overline{Y})^2 + \sum_i (\hat{Y}_i - \overline{Y})^2 - 2 \cdot b_1^{\,2} \sum_i (X_i - \overline{X})^2$$

En utilisant la *formule n° 2* :

$$\sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - \overline{Y})^2 + \sum_i (\hat{Y}_i - \overline{Y})^2 - 2 \sum_i (\hat{Y}_i - \overline{Y})^2$$

On aboutit enfin à l'égalité fondamentale :

$$\sum_i (Y_i - \overline{Y})^2 = \sum_i (\hat{Y}_i - \overline{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

La SCE (Somme des Carrés des Ecarts) totale est égale à la somme des carrés des écarts dus au modèle augmentée de la somme des carrés des écarts dus aux erreurs

$$\text{SCE totale} = \text{SCE modèle} + \text{SCE erreur}.$$

Cette formule montre que :

Les variations de Y autour de sa moyenne, c'est-à-dire **SCE Totale** (**SS Total** pour Sum of Squares en anglais) peuvent être expliquées par :

- le modèle grâce à **SCE Modèle** (**SS Model** en anglais) ;
- et ce qui ne peut être expliqué par le modèle, est contenu dans **SCE Erreur** (**SS Error** en anglais).

L'erreur est aussi appelée le « **résidu** ».

Standard de présentation de l'Analyse de la Variance

On a l'habitude de représenter l'analyse de la variance sous forme d'un tableau, faisant apparaître les 3 sources de variation : le total en 3ième ligne qui se décompose en la partie modèle et la partie erreur.
A chaque source de variation (Total, Modèle, Erreur) correspond un nombre de degrés de liberté (ddl) respectivement égal à n-1, p, n-p-1,

n : nombre d'observations
p : nombre de variables régresseurs (la variable X_0 , constante égale à 1, correspondant au paramètre β_0, n'est pas comprise).

Nous présentons le tableau général de l'analyse de variance pour p régresseurs. Pour la régression simple, p=1 (une seule variable régresseur).

Tableau 1. 3 Analyse de variance (version anglaise)

Source	DF	Sum of Squares	Mean Square
MODEL	p	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 / p$
ERROR	n-p-1	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 /(n-p-1)$
TOTAL	n-1	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	

Abréviations:

DF : Degrees of Freedom se traduit par degrés de liberté (ddl).
Ils vérifient : $DF_{total}=DF_{model} + DF_{erreur}$
SS : Sum of Squares se traduit par Somme des Carrés des Ecarts (SCE)
MS : Mean Square, est le rapport SS/DF, relatif soit au modèle soit à l'erreur
MSE : Mean Square Error = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 /(n-p-1)$ représente le carré de l'écart moyen résiduel.

Tous ces « indicateurs » SS, MS, MSE, vont jouer un rôle important dans l'appréciation du modèle calculé à partir des observations.

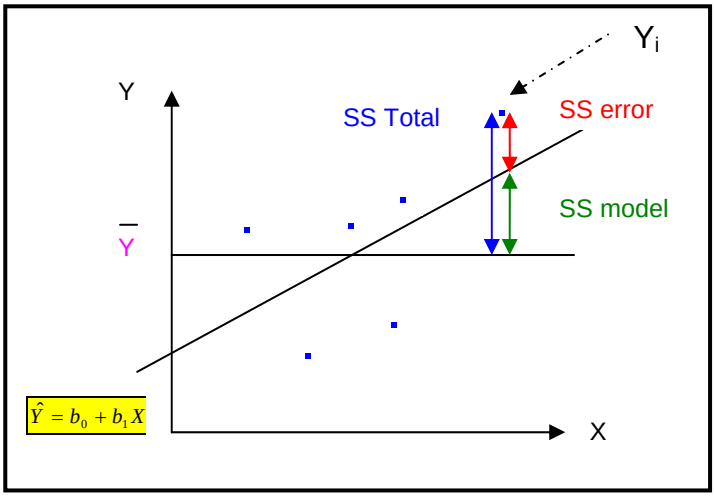


Figure 1.5 Décomposition des SS –Sums of Squares

La figure 1.5 montre les liens entre **SS total**, **SS model** et **SS error** lorsque l'on somme les carrés des écarts sur tous les points i.

Il est remarquable que la formule de décomposition de l'écart total en un point i, vu au § 1.2.4.

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

prennela même forme pour la somme des carrés.

$$\sum_i (Y_i - \overline{Y})^2 = \sum_i (\hat{Y}_i - \overline{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

Comment apprécier globalement la régression

Les deux quantités SCE totale (SS total) et SCE modèle (SS model) sont des sommes de carrés donc toujours positives ou nulles et telles que
$\text{SCE Modèle} \leq \text{SCE Totale}$.

Le rapport $\dfrac{\text{SCE Modèle}}{\text{SCE Totale}}$ est donc compris entre 0 et 1.

On appelle ce rapport **le coefficient de détermination**

$$R^2 = \frac{\text{SCE Modèle}}{\text{SCE Totale}} = \frac{\text{SS model}}{\text{SS Total}}$$

Cas particulier :
Si tous les points Y_i observés sont alignés sur la droite de régression, le modèle est parfaitement adapté et SCE Erreur = 0,

Dans ce cas: $\dfrac{\text{SCE Modèle}}{\text{SCE Totale}} = 1$

Interprétation de R^2

R^2 qui varie entre 0 et 1, mesure la proportion de variation totale de Y autour de la moyenne expliquée par la régression, c'est-à-dire prise en compte par le modèle. Plus R^2 se rapproche de la valeur 1, meilleure est l'adéquation du modèle aux données.

Un R^2 faible signifie que le modèle a un faible pouvoir explicatif.
On démontre que R^2 représente aussi le carré du **coefficient de corrélation linéaire** entre Y et Y estimé:

$$R^2 = \text{Corr}^2(Y, \hat{Y})$$

Dans le cas de la régression simple, R est aussi la valeur absolue du coefficient de corrélation linéaire entre Y et X.

$$R = |\text{Corr}\ (Y, X)|$$

Lien entre coefficient de corrélation de 2 variables et le cosinus de leur angle

Soient 2 vecteurs X_1 et X_2 définis dans un espace R^n (espace des n observations), le coefficient de corrélation entre X_1 et X_2 est aussi le cosinus de l'angle θ entre ces 2 vecteurs.

En utilisant les conventions de notation, le produit scalaire de 2 vecteurs X_1 et X_2 se note $< X_1, X_2 > = \|X_1\| * \|X_2\| \text{Cos}(\theta)$

On a :

$$\text{Cos}(X_1, X_2) = \frac{< X_1, X_2 >}{(< X_1, X_1 >< X_2, X_2 >)^{1/2}} = \frac{1}{n} * \sum_{i=1,n} \frac{(X_{1,i} - \overline{X})(X_{2,i} - \overline{X})}{s_1 * s_2} = \text{Corrélation } (X_1, X_2)$$

$s_1 * s_2$ étant le produit des écarts-type des 2 vecteurs.

L'interprétation d'un coefficient de corrélation comme un cosinus est une propriété importante.

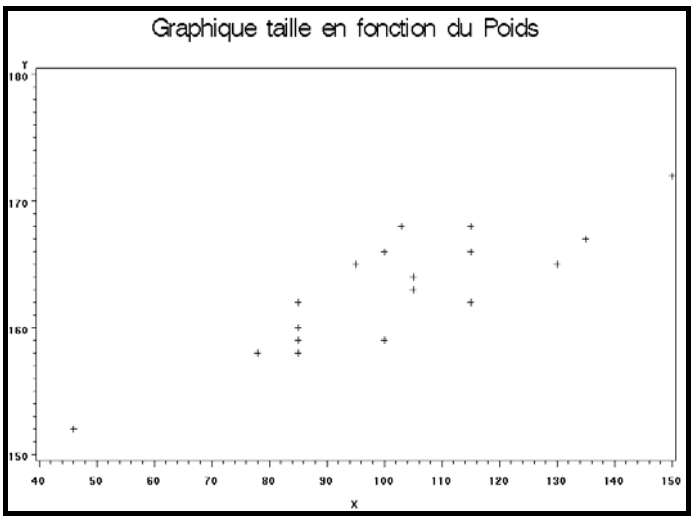
Comme le remarque TOMASSONE (1992), les variables X n'étant pas des variables aléatoires, il est plus correct de parler de « *cosinus* » des angles formés par les vecteurs associés, en réservant le terme « *coefficient de régression* » pour sa similitude avec l'estimation de ce coefficient à partir d'un échantillon.

Exemple : Régression de la Taille en fonction du Poids

Sur les données du tableau 1.2**,** la première étape consiste à « regarder » les données pour vérifier qu'une liaison linéaire est envisageable (Proc GPLOT). Puis en deuxième étape on calcule le coefficient de corrélation (Proc CORR). *Cette deuxième étape non indispensable en régression simple deviendra essentielle en régression multiple*. Enfin on effectue une régression linéaire (Proc REG).

Programme SAS

```
Proc gplot data=libreg.tailpoid;
 plot Y*X;
 title ' Graphique taille en fonction du
Poids ';
Proc corr data=libreg.tailpoid;
title 'Corrélation ';
var Y X;
Proc REG data=libreg.tailpoid;
title 'Régression de la Taille en fonction du
Poids ';
model y=x;
run;
```



Corrélation

The CORR Procedure

2 Variables: Y X

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Y	20	163.25000	4.58688	3265	152.00000	172.00000
X	20	101.35000	22.56284	2027	46.00000	150.00000

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	Y	X
Y	1.00000	0.83771 <.0001
X	0.83771 <.0001	1.00000

Le coefficient de corrélation CORR entre Y et X vaut 0.83771.

Sortie standard de la Proc REG sans options

Régression de la Taille en fonction du Poids

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	280.52918	280.52918	42.35	<.0001
Error	18	119.22082	6.62338		
Corrected Total	19	399.75000			

Root MSE	2.57359	R-Square	0.7018
Dependent Mean	163.25000	Adj R-Sq	0.6852
Coeff Var	1.57647		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	145.98994	2.71384	53.79	<.0001
X	1	0.17030	0.02617	6.51	<.0001

Dans la sortie de Proc REG on obtient d'abord le tableau d'analyse de la variance, puis les estimations des paramètres.

Lecture de l'Analyse de la Variance

SS Model = $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ = 280.52918

SS Error = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ = 119.22082

SS Total = = 399.75

Mean Square Model = $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 / p$ = 280.52918

Mean Square Error = $\boxed{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 /(n-p-1)}$ = 6.62338

Root MSE = $\sqrt{\text{MS ERROR}}$ = 2.57359

Dependant Mean = $\overline{Y} = 163.25$

R-Square = $\boxed{\frac{\text{SS Model}}{\text{SS Total}}}$ = 0.7018 = CORR(X,Y)^2 = (0.83771)^2.

Autres indicateurs

• **CV =1.57647 C'est le Coefficient de Variation =** $\boxed{\frac{\text{Root MSE}}{\text{Dep Mean}} * 100}$

Le CV est un indicateur sans dimension -exprimé en %- permettant de comparer l'écart moyen résiduel à la moyenne de la variable dépendante Y. Ce pourcentage est plutôt utilisé pour comparer 2 modèles (donc 2 CV) portant sur le même jeu de données.

• **Le coefficient R^2 ajusté , Adj R-sq**

Le R^2 ajusté (utilisé en régression multiple) tient compte du nombre de paramètres du modèle.

$$\boxed{R^2 \text{ajusté} = 1 - \frac{(n - \text{intercept})(1 - R^2)}{n - p}}$$

Avec *Intercept*=0, si il n'y a pas de constante b_0 à l'origine[7] sinon *Intercept* =1.

Le reproche fait au coefficient de détermination est qu'il peut approcher la valeur 1, interprété comme un ajustement parfait, si on ajoute suffisamment de variables régresseurs.

Le R^2 ajusté tient compte du rapport p/n entre le nombre de paramètres du modèle et le nombre d'observations.

Selon certains auteurs ce coefficient permet de comparer des modèles de régression sur différents ensembles de données, mais il ne fait pas l'unanimité.

Attention : Adj R-sq peut prendre des valeurs inférieures à zéro !

[7] S'il n'y a pas de constante b_0 à l'origine, les statistiques relatives à l'analyse de la variance n'ont pas la même interprétation.

Lecture du tableau des paramètres

Intercept = $b_0 =$ 145.98 donne la valeur de la constante à l'origine. On peut remarquer que dans cet exemple, cette valeur n'a pas de signification dans le monde physique. On ne peut concevoir qu'à un poids de valeur nulle corresponde une taille de 145.98.

La pente de la droite (coefficient de X) = $b_1 =$ 0.1703.

On l'interprète comme augmentation de la taille lorsque le poids augmente de une unité.

Equation de la droite : $\boxed{\text{Taille} = 145.98 + 0.1703 * \text{Poids}}$

Là encore il faut se préserver de toute interprétation causale. Peut-on agir et augmenter le poids en espérant faire augmenter la taille ?

Nous verrons les autres indicateurs dans la suite du chapitre.

Pour mieux comprendre la technique de la régression, voyons certaines représentations géométriques.

1.2.6. Représentations géométriques

Régression simple de Y sur X

Afin d'avoir une idée géométrique de la régression prenons un exemple avec n=3 observations (y_1,x_1), (y_2,x_2) et (y_3,x_3).

Le vecteur réponse Y = (y_1,y_2,y_3), et le vecteur régresseur X = (x_1,x_2,x_3) peuvent se représenter dans l'espace à 3 dimensions des observations. On nomme 1,2,3 les axes de ce repère.
Dans l'espace des observations représenté figure 1.6 [8], la droite Δ des constantes a pour vecteur directeur (1,1,1).

[8] La figure 1.6 est une synthèse des graphiques de DRAPER & SMITH (1966) pp112-113 et SAPORTA (2006) p208

Q est la projection orthogonale de Y sur Δ de coordonnées $(\overline{Y},\overline{Y},\overline{Y})$

P est la projection orthogonale de Y sur le plan (X,Δ) et représente $\hat{Y} = (\hat{Y}_1, \hat{Y}_2,$

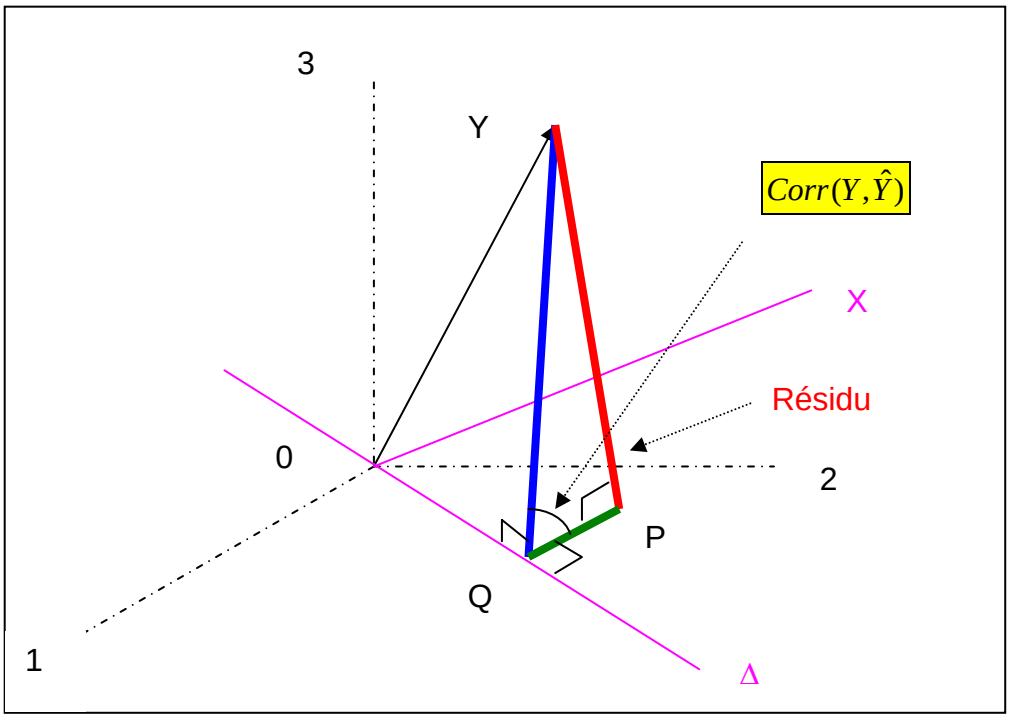


Figure 1.6 Régression de Y sur (X, Δ) dans l'espace des 3 observations.

L'interprétation géométrique de la régression est la suivante :

Régresser Y sur (Δ et X) consiste à **projeter orthogonalement** Y sur le plan (Δ,X) ce qui donne le point P.
Si d'autre part on projette Y sur la droite Δ, on obtient le point Q.
Par le théorème des 3 droites perpendiculaires Q est aussi la projection orthogonale de P sur Δ.

Dans le triangle YQP, rectangle en P, on peut appliquer le théorème de Pythagore :

$$YQ^2 = YP^2 + PQ^2$$

$$\sum_i (Y_i - \overline{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \overline{Y})^2$$

La longueur YQ^2 représente la somme des carrés corrigée **SCE Totale (SS Total)**.
La longueur YP^2 représente la somme des carrés non expliquée par le régression **SCE Erreur (SS error)** .
La longueur PQ^2 représente la somme des carrés expliquée par la régression soit **SCE Modèle (SS model)**.

C'est l'équation fondamentale de l'analyse de la variance vue précédemment :

$$\text{SS Total} = \text{SS Model} + \text{SS Error}$$

Le coefficient de détermination R^2 est le rapport $\frac{PQ^2}{YQ^2}$.

R^2 représente donc le carré du cosinus de l'angle (YQ, QP), c'est à dire l'angle entre Y et $\hat{Y}$.

Plus l'angle entre Y et $\hat{Y}$ est faible, meilleur est le *pouvoir explicatif du modèle*.

Et maintenant il suffit de généraliser mentalement à l'ordre n cette représentation à 3 dimensions.

Remarque en régression multiple
Si au lieu d'avoir une seule variable régresseur X, on avait plusieurs variables $X_1,\ldots X_P$, alors le plan de projection (X,Δ) serait remplacé par l'hyperplan formé par les vecteurs $X_1,\ldots X_p$, Δ.
Régresser Y sur les p variables régresseurs consisterait à projeter orthogonalement Y sur l'hyperplan déterminé par $X_1,\ldots X_p$, Δ.

Distribution en un point fixé de X

Jusqu'ici, on ne s'est appuyé que sur des calculs algébriques et sur des notions de géométrie, sans faire appel à des notions de statistique. On ne cherchait que la droite d'ajustement sur l'échantillon.
Aucune supposition n'a été nécessaire dans toutes les démonstrations.

Si maintenant, on souhaite utiliser les résultats obtenus à partir des observations, vues comme un échantillon, pour **inférer** sur la population, il faut faire appel à des notions de probabilité, et de statistique puisque dans les relevés de données (exemple : *Poids et Taille*) à notre disposition on n'a qu'un échantillon de valeurs et non toute la population.

Sur la figure 1.7, on remarque que pour une même valeur du *Poids*, par exemple 85, il y a plusieurs valeurs possibles de la *Taille* (158, 159, 160 et 162).

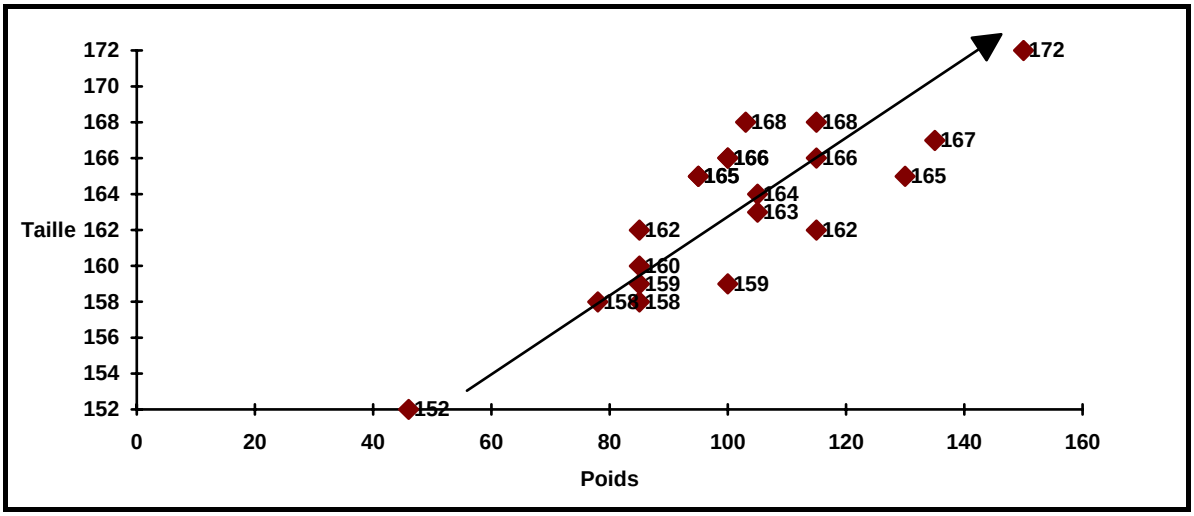


Figure 1.7 Taille en fonction du Poids

Il n'y a pas de valeur unique associée à une valeur X_i mais une distribution de valeurs.

Pour chaque valeur du poids (X) existe une distribution théorique des tailles (Y). Les valeurs de centrage sont les espérances des tailles de la population correspondant à chaque poids X_i. L'espérance (moyenne théorique μ_i) de chaque distribution de Y, est appelée « statistiquement parlant » l'*espérance de Y_i sachant X_i* que l'on note $E(Y_i/X_i)$.
L'hypothèse de la régression linéaire est que les μ_i sont alignés sur la vraie droite de régression qui est inconnue.

Remarque : pour simplifier l'écriture on note $E(Y_i)$ au lieu de $E(Y_i/X_i)$, soit :

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 X_i$$

Représentation de X fixé et Y aléatoire

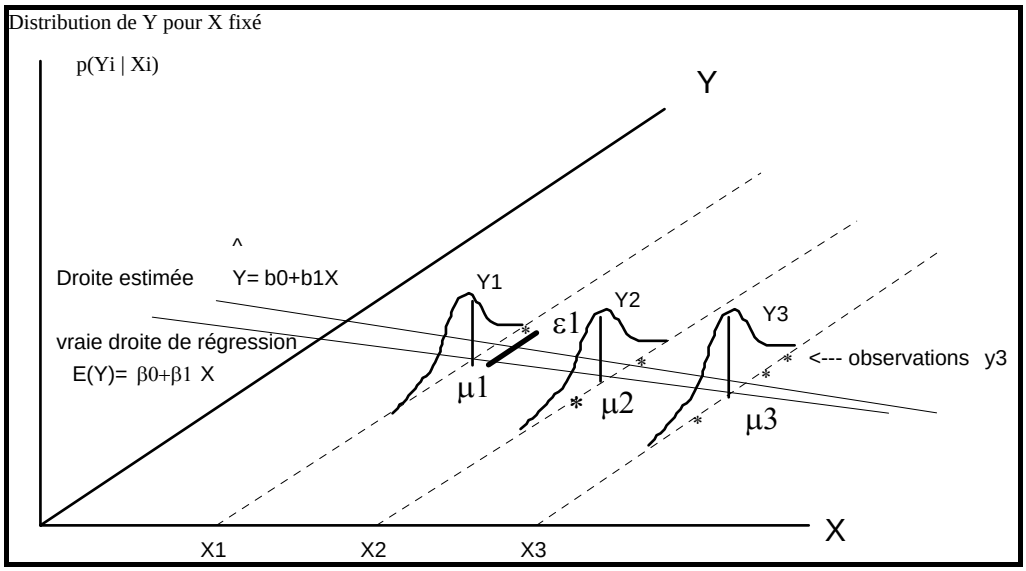


Figure1.8 Distributions de Y pour X fixé

Pour un même poids X_1 fixé on a une distribution de taille Y_1, dont on a observé une réalisation y1, ou plusieurs.

Par exemple sur le graphique Taille*Poids de la figure 1.7, on remarque que pour X=46 on a une seule valeur observée Y=152, tandis que pour X=85 on observe plusieurs valeurs de Y (158, 159, 160 et 162).

Chaque Y_i est une variable aléatoire qui a une distribution de probabilité de Y_i sachant X_i notée $p(Y_i | X_i)$. Des hypothèses sur la régularité de ces distributions devront être faites :

- les **distributions**, pour tous les points X_i, sont supposées **normales**
- les **espérances** des distributions sont centrées sur la droite de régression
- les **variances** de chaque Y_i conditionnellement à X_i sont toutes **égales**.

De plus les variables aléatoires Y_i ne doivent pas être reliées entre elles, elles sont supposées **indépendantes**.
Ces **suppositions** se résument ainsi :

Les variables aléatoires Y_i sont indépendantes,

d'espérance et de variance :

$$E(Y_i) = \beta_0 + \beta_1 X$$

$$\text{variance}(Y_i) = \sigma^2$$

Il faut avoir à l'esprit que $E(Y_i)$ est une espérance conditionnelle. De même lorsque l'on parle de variance de Y, c'est sous-entendu, variance conditionnellement à X.

« Vraie » droite de régression et droite estimée par la régression

La figure1.8 montrant les distributions de Y pour X fixé est une illustration du modèle de régression linéaire.
Toujours en supposant que le modèle linéaire postulé est le véritable modèle, on obtiendrait le vraie droite de régression $E(Y) = \beta_0 + \beta_1 X$, si on avait à notre disposition toute la population.
Comme on n'a qu'un échantillon d'observations, on n'a qu'une **estimation** $\hat{Y} = b_0 + b_1 X$ ou droite estimée par la régression.

A propos des erreurs

L'erreur théorique ε_i représente l'écart entre Y_i observé et l'espérance $E(Y_i)$ non observable. On notera que ε_i non plus n'est pas observable. Ce qui est observable c'est l'erreur e_i correspondant à l'écart entre Y_i observé et $\hat{Y}_i$, son estimation par le modèle.

Le résidu observé e_i est une **estimation** de l'erreur inobservable ε_i .

1.3. Glissement fonctionnel de la méthode des Moindres Carrés Ordinaires à la Régression.

De la théorie des erreurs en astronomie à l'étude des moyennes en sciences sociales, un siècle les sépare.

1.3.1. De l'Astronomie...

Historiquement la méthode des moindres carrés à d'abord été développée par LEGENDRE en 1805, pour répondre à une question posée par les astronomes et les spécialistes de la géodésie comme le rapporte DESROSIÈRES (1993) :

"Comment combiner des observations effectuées dans des conditions différentes, afin d'obtenir les meilleures estimations possibles de plusieurs grandeurs astronomiques ou terrestres liées entre elles par une relation linéaire?".

Ces grandeurs sont mesurées par des instruments imparfaits, et par des observateurs qui ne sont pas tous identiques. **Il y a des erreurs de mesures dans les observations**.

De là provient le vocabulaire : observation, écart, erreur ou résidu.
Vous pouvez trouver sur internet une traduction anglaise de ce premier article scientifique de Legendre sur les moindres carrés (Least Squares)
http://www.stat.ucla.edu/history/legendre.pdf.

1.3.2. … Aux Sciences Sociales

En s'appuyant sur :

1. Le **théorème central limite** (LAPLACE 1810) montrant que même si la distribution de probabilité des erreurs ne suit pas une loi normale, celle de la **moyenne** tend vers une loi normale, quand le nombre des observations s'accroît indéfiniment,

2. La synthèse opérée par Laplace et Gauss vers 1810 entre:

   - comment combiner au mieux des observations imparfaites ?
   Réponse : en utilisant le milieu (la moyenne),
   - comment estimer le degré de confiance que mérite une estimation ?
   Réponse : en terme de probabilité,

Galton inventeur de la "régression" et PEARSON inventeur de la "corrélation" appliquèrent l'ajustement des moindres carrés à des données sociales dans les années 1880.

Nous reproduisons ci-après le graphique[9] de GALTON, révélateur d'une « *Reversion* », et accessible sur internet :
http://www.stat.ucla.edu/history/regression.gif .

1.3.3. Galton Diagram Regression

En 1885 GALTON réalisa le tableau qui croise la taille de 928 enfants (devenus adultes) nés de 203 parents, en fonction de la taille moyenne de leurs parents (la taille de la mère étant préalablement multipliée par un coefficient 1.08).

[9] F GALTON, Regression towards mediocrity in hereditary stature", *Journal of the Anthropological Institute* **15** (1886), 246-263.

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

Height of the mid-parent in inches	Height of the adult child														Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7			
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1886a).

Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties); he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

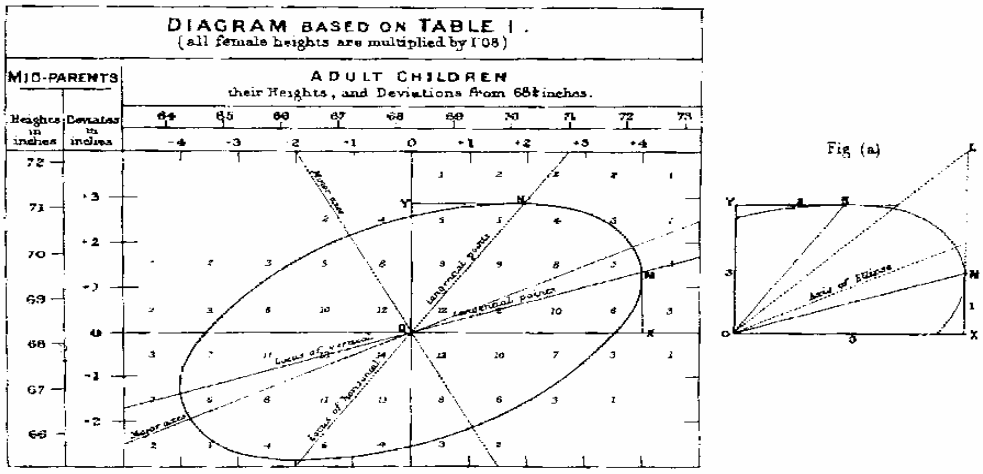


Figure 8.7. Galton's smoothed tabulation of Table 8.1, with one of the "concentric and similar ellipses" drawn in. The geometric relationship of the two regression lines to the ellipse is also shown. (From Galton, 1886a.)

En présentant ce tableau sous forme d'un graphique, GALTON remarqua que l'on pouvait voir des ellipses de densités. Si les parents sont plus grands que la moyenne, les enfants seront également plus grands que la moyenne mais avec une taille plus proche de la moyenne que celle de leurs parents.

Si les parents sont de petites tailles, leurs enfants seront également plus petits que la moyenne, mais avec une taille plus proche de la moyenne que celle de leurs parents. Il y a régression vers la moyenne.

D'où le terme de « régression ».

Ce n'est que vers les années 1930 que le formalisme de la méthode des moindres carrés associé à une interprétation probabiliste est devenu la « **Régression** » (ARMATTE (1995)).

Le glissement des méthodes d'analyse, des erreurs en Astronomie vers des estimations de moyennes en Sciences Sociales, a conduit à appeler **erreur** ou **perturbation** ou encore **aléa,** l'écart de Y par rapport à sa moyenne.
Le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

où les erreurs ε_i sont des aléas indépendants d'espérance =0 de variance σ^2

1.3.4. Formalisation des Suppositions

L'ensemble des suppositions nécessaires pour élaborer les tests statistiques se résume ainsi:

- l' erreur ε_i est une variable aléatoire d'espérance nulle et de variance constante σ^2.

$$E(\varepsilon_i) = 0 \text{ et } Var(\varepsilon_i) = \sigma^2$$

- l' erreur ε_i est non corrélée à ε_j.

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ pour } i \neq j$$

- les erreurs ε_i sont normalement distribuées

$$\varepsilon_i \approx N(0, \sigma^2)$$

On résume souvent ces 3 suppositions par l'expression "***i i d selon une loi normale***" qui signifie

Indépendantes et **I**dentiquement **D**istribuées selon une loi normale

Il faut de plus que les variables aléatoires Y conditionnellement à X soient indépendantes, et que les régresseurs X_j soient non aléatoires et non corrélés.

Lorsque ces suppositions sont vérifiées, l'estimateur MCO est non biaisé et efficace (de variance minimum). En anglais on utilise l'acronyme BLUE (BEST Linear Unbiased Estimator)

Nous verrons au paragraphe suivant, comment interviennent ces suppositions sur les aléas dans les raisonnements statistiques.

1.4. Confiance à accorder aux résultats

Pour inférer de l'échantillon à la population dont sont issues les observations, la logique statistique nous conduit à effectuer des tests d'hypothèses, et à déterminer des intervalles de confiance autour des valeurs estimées.

Successivement on va chercher à :

- **tester** la signification globale de la régression,
- **tester** l'hypothèse nulle β_1=0 et à **calculer** l'intervalle de confiance de β_1,
- **tester** l'hypothèse nulle β_0=0 et à **calculer** l'intervalle de confiance de β_0,
- **calculer** la précision de l'estimation de Y pour la moyenne et pour une observation individuelle.

1.4.1. Test de la signification globale de la régression

Ce test a surtout un intérêt dans le cadre de la régresssion multiple, c'est à dire avec p régresseurs. En anticipant sur le chapitre 2, qui présente la régression multiple, on généralise le modèle de régression à un régresseur au cas d'un modèle à p régresseurs $X_1, X_2 \cdots X_p$:

$$\boxed{\mu_i = E(Y_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

Ce test permet de connaître l'apport global de l'ensemble des variables $X_{1,},...,X_p$ à la détermination de Y.
On veut tester l'hypothèse nulle:
H_0: $\beta_1 = \cdots = \beta_p = 0$ contre
H_a: Il existe au moins un β_j parmi $\beta_1, \cdots, \beta_p$ non égal à 0.

On calcule la statistique de test $\boxed{F = \dfrac{\text{MS model}}{\text{MS error}}}$

avec $\boxed{\text{MSmodel} = \dfrac{\text{SS model}}{p}}$ et $\boxed{\text{MSerror} = \dfrac{\text{SS error}}{n-p-1}}$ représentant respectivement une somme de carrés des écarts moyens respectivement pour le modèle et pour l'erreur.

Si H_0 est vraie et sous réserve des suppositions suivantes, ce rapport F est une valeur observée d'une variable qui suit une loi de Fisher-Snedecor à p et n-p-1 degrés de liberté.

Si les ε_i sont indépendants et suivent une loi normale de même variance
$\varepsilon_i \approx N(0, \sigma^2)$

Alors la statistique F suit une loi de Fisher-Snedecor

$$\boxed{F = \frac{\text{MS model}}{\text{MS error}} \approx F(p, n-p-1)}$$

Règle de décision

Si F observé $\geq F_{1-\alpha}(p, n-p-1)$
Alors H_0: $\beta_1 = \cdots = \beta_p = 0$ doit être rejetée au niveau α

où $F_{1-\alpha}(p, n-p-1)$ représente le quantile d'ordre $(1-\alpha)$ de la loi de Fisher-Snedecor à (p) et (n-p-1) degrés de liberté.

Note: Dans SAS, la fonction de répartition inverse pour une loi de Fisher-Snedecor est donnée par la fonction FINV.
Instruction SAS ⇨ $\boxed{F = \text{FINV}(1-\alpha, p, n-p-1)}$
Avec n= nombre d'observations et p = nombre de régresseurs (non compris la constante).

Pour éviter de raisonner sur F, SAS fournit la *p-value* associée au F observé. La *p-value* est le niveau de significativité du test de Fisher-Snedecor, c'est-à-dire la probabilité de dépasser le F observé si l'hypothèse nulle est vraie.
On compare la *p-value* au risque α choisi (par exemple α=0.05).

Raisonnement sur la p-value

Si p-value $\leq \alpha$
Alors on rejette l'hypothèse nulle $\beta_1 = \cdots = \beta_p = 0$

Interprétation
On dit que la régression est significative au niveau α.
Le modèle retenu améliore la prévision de Y par rapport à la simple moyenne des Y.
Pour la régresssion simple, ce test porte uniquement sur le paramètre β_1.

Ce test fournit un moyen d'apprécier la régression dans son ensemble, ce qui ne signifie pas que chacun des coefficients de la régression soit significativement différent de 0.

1.4.2. Statistiques liées au paramètre β_1

Pour s'assurer de la significativité du paramètre β_1, on va dans une première étape calculer la variance de b_1, puis en deuxième étape tester l'hypothèse nulle β_1=0, en troisième étape on pourra alors déterminer un intervalle de confiance pour β_1 autour de b_1.

Calcul de la variance de b_1

On a vu que b_1 est le rapport de la covariance entre X et Y divisé par la variance de X :

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

On développe le numérateur

$$b_1 = \frac{\sum (X_i - \overline{X})Y_i - \overline{Y}\sum (X_i - \overline{X})}{\sum (X_i - \overline{X})^2}$$

Comme le 2ième terme du numérateur est nul, par définition de la moyenne $\sum (X_i - \overline{X})$ =0, il ne reste que le 1er terme.

$$b_1 = \frac{((X_1 - \overline{X})Y_1 + \cdots + (X_n - \overline{X})Y_n)}{\sum (X_i - \overline{X})^2}$$

On ne peut calculer la variance de b_1 que si on fait des suppositions sur les X_i et sur les liaisons entre les Y_i.

Suppositions pour calculer la variance de b_1

Si les X_i sont non aléatoires
Si les Y_i sont non corrélés et de même variance σ^2
Et comme par construction $\mathrm{Cov}(\overline{Y}, b_1) = 0$
Alors :

$$\mathrm{Var}(b_1) = a_1{}^2\mathrm{Var}(Y_1) + \cdots + a_n{}^2\mathrm{Var}(Y_n)$$

avec $a_i = \frac{(X_i - \overline{X})}{\sum (X_i - \overline{X})^2}$ assimilés à des constantes

Ce qui permet d'aboutir à : $\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$

σ^2 représente la variance **inconnue** de Y. Il faut de nouveau faire une supposition.

Supposition

Si le modèle postulé est le modèle correct Alors σ^2 peut être estimé par les erreurs entre les Y_i observés et $\hat{Y}_i$

Mean Square Error= $\boxed{MSE = s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}}$

Note : Pour la régression multiple : $\boxed{MSE = s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-p-1}}$

Compte tenu de toutes ces suppositions, l'estimateur de l'écart-type de b_1 devient :

$$\boxed{s(b_1) = \frac{s}{\sqrt{\sum (X_i - \overline{X})^2}}}$$

Remarques:

- La variance de b_1 est inversement proportionnelle à la dispersion des X_i autour de la moyenne.
  Donc, si on veut améliorer la précision de b_1 il faut, si possible, augmenter la variance empirique des X_i.
- La variance de b_1 est inversement proportionnelle à (n-2), n étant la taille de l'échantillon.
  Donc, si on veut améliorer la précision de b_1 il faut augmenter la taille de l'échantillon.

Test portant sur le paramètre β_1

On s'intéresse au test de l'hypothèse nulle:
H_0 : paramètre $\beta_1 = 0$ contre H_a : paramètre $\beta_1 \neq 0$

On calcule la statistique de test $\boxed{Tobservé = \frac{b_1}{s(b_1)}}$

Si $\beta_1=0$ la statistique Tobservé suit une loi de Student, **sous l'hypothèse** que les erreurs soient indépendantes et identiquement distribuées selon la loi Normale.

Suppositions

Si $\beta_1 = 0$
Si $\varepsilon_i \approx N(0,\sigma^2)$
Alors T observé suit une loi de Student à n-1 degrés de liberté

Raisonnement

On compare la *p-value* associée à T observé, au risque α choisi (par ex:α=0.05).

Si p-value $\leq \alpha$
Alors on rejette l'hypothèse $\beta_1 = 0$

Conclusion *:* β_1 est significativement différent de zéro au niveau α

Calcul de l'intervalle de confiance de β_1

On peut calculer un intervalle de confiance (IC de niveau 1-α) autour de b_1, ce qui permet de statuer sur le paramètre β_1 :

$$IC_{1-\alpha}(\beta_1)=\left[b_1 - t_{1-\alpha/2}\cdot s(b_1);b_1 + t_{1-\alpha/2}\cdot s(b_1)\right]$$

où $t_{1-\alpha/2}$ représente le quantile d'ordre 1-α/2 de la loi de Student à (n-2) degrés de liberté.

Note

- Dans SAS, la fonction de répartition inverse pour une loi de Student est donnée par la fonction TINV.
  Instruction SAS ⇨ $T = TINV(1-\alpha/2, n-2)$ avec n= nombre d'observations
- Dans le cas de la régression multiple avec p=nombre de régresseurs, la formule précédente devient:
  Instruction SAS ⇨ $T = TINV(1-\alpha/2, n-p-1)$

En pratique : si α=5% et si n est assez grand (n>30), pour approcher la loi de Student par la loi Normale,

Alors $$IC_{0.95}(\beta_1)=\left[b_1 - 1.96\cdot s(b_1);b_1 + 1.96\cdot s(b_1)\right]$$

Interprétation

Si la valeur 0 est dans l'intervalle de confiance de β_1, alors l'introduction de la variable X dans le modèle n'apporte aucun pouvoir explicatif sur Y.

1.4.3. Statistiques liées au paramètre β_0

La première étape consiste à calculer la variance de b_0, puis en deuxième étape à tester l'hypothèse nulle β_0=0, en troisième étape on pourra alors déterminer un intervalle de confiance pour β_0.

Calcul de la variance de b_0

On a vu que b_0 vaut :

$$b_0 = \overline{Y} - b_1\overline{X}$$

la variance vaut:

$$Var(b_0) = Var(\overline{Y} - b_1\overline{X})$$

Raisonnement pour calculer la variance de b_0

Pour pouvoir calculer la variance il faut faire des *suppositions* sur les termes de cette expression.
On suppose que les X_i sont non aléatoires.

Seuls la moyenne des Y_i et le coefficient b_1 sont des variables aléatoires. On peut montrer de plus que la covariance entre $\overline{Y}$ et le coefficient b_1 est nulle[10] .

Suppositions pour calculer la variance de b_0

Si les X_i sont non aléatoires
Si les Yi sont non corrélés et de même variance σ^2
Et comme par construction $Cov(\overline{Y}, b_1) = 0$

Alors :

$$Var(b_0) = Var(\overline{Y}) + \overline{X}^2 Var(b_1) =$$
$$= \frac{\sigma^2}{n} + \overline{X}^2 \frac{\sigma^2}{\sum(X_i - \overline{X})^2} = \frac{\sigma^2 \sum X_i^2}{n\sum(X_i - \overline{X})^2}$$

σ^2 représente la variance **inconnue** de Y. Il faut de nouveau faire une supposition.

Supposition

Si le modèle postulé est le modèle correct
Alors σ^2 peut être estimé par les erreurs entre les Y observés et $\hat{Y}$

$$s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = MSE$$

L'estimateur de la variance de b_0 devient :

$$s^2(b_0) = \frac{s^2 \sum X_i^2}{n\sum(X_i - \overline{X})^2}$$

Remarque:
La variance de b_0 est proportionnelle à la somme des carrés des X_i.
Si le plan d'expérience est tel que les valeurs des X_i sont très grandes, la variance de b_0 sera très grande, et l'estimation de b_0 n'aura aucune signification.

Test portant sur le paramètre β_0

Test de l'hypothèse nulle H_0 : paramètre $\beta_0 = 0$ contre H_a : paramètre $\beta_0 \neq 0$

On calcule la statistique de test $Tobservé = \frac{b_0}{s(b_0)}$

[10] voir démonstration dans NETER, WASSERMAN, KUTNER pp75-77.

Si β_0=0 la statistique Tobservé suit une loi de Student à n-2 degrés de liberté, **sous l'hypothèse** que les erreurs sont indépendantes et identiquement distribuées selon la loi Normale.

Supposition

$$\text{Si } \varepsilon_i \approx N(0,\sigma^2)$$
Alors T observé suit une loi de Student

Raisonnement

On compare la *p-value* associée à T observé, c'est-à-dire la probabilité de dépasser le T observé en valeur absolue, au risque α choisi (par exemple α=0.05).

Si p-value $\leq \alpha$
Alors on rejette l'hypothèse β_0 =0

Conclusion **β_0 est significativement différent de zéro au niveau α**

Calcul de l'intervalle de confiance de β_0

On peut assigner un intervalle de confiance autour de b_0, ce qui permet de statuer sur le paramètre β_0:

$$IC_{1-\alpha}(\beta_0)=[b_0 - t_{1-\alpha/2}\cdot s(b_0);b_0 + t_{1-\alpha/2}\cdot s(b_0)]$$

où $t_{1-\alpha/2}$ représente le quantile d'ordre 1-α/2 de la loi de Student à n-2 degrés de liberté.

Note

- Dans SAS, la fonction de répartition inverse pour une loi de Student est donnée par la fonction TINV.
  Instruction SAS ⇨ $T = TINV(1-\alpha/2, n-2)$ avec n= nombre d'observations

- Dans le cas de la régression multiple avec p=nombre de régresseurs, la formule précédente devient:
  Instruction SAS ⇨ $T = TINV(1-\alpha/2, n-p-1)$

En pratique si on choisit le risque α=5% et si n est assez grand (n>30) pour approcher la loi de Student par la loi Normale, alors l'intervalle de confiance de β_0 à 95% est donné par :

$$IC_{0.95}(\beta_0)=[b_0 - 1.96\cdot s(b_0);b_0 + 1.96\cdot s(b_0)]$$

Interprétation

Si la valeur 0 est dans l'intervalle de confiance de β_0, alors la droite de régression passe par l'origine.

Exemple d'estimation des paramètres avec Proc REG

Sur l'exemple de la Taille en fonction du Poids

Programme SAS

```
Proc REG data=libreg.tailpoid outest=TableSortie;
title 'Régression de la Taille en fonction du Poids ';
model y=x ;
proc Print;title "Table de l'option outest";
run;
```

Sortie de Proc REG

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	280.52918	280.52918	42.35	<.0001
Error	18	119.22082	6.62338		
Corrected Total	19	399.75000			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	145.98994	2.71384	53.79	<.0001
X	1	0.17030	0.02617	6.51	<.0001

Table de l'option outest

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	X	Y
1	MODEL1	PARMS	Y	2.57359	145.990	0.17030	-1

Interprétation du test de la signification globale de la régression

La statistique $F = \frac{\text{MS model}}{\text{MS error}} = \frac{280.529}{6.62} = 42.35$ indique que globalement le modèle avec le régresseur Poids améliore la prévision de la Taille, par rapport à la moyenne seule dans le modèle.

Interprétation des estimations des paramètres

L'estimateur de β_0 a pour valeur 145.98994. Son écart type vaut 2.71384.

La statistique de Test $\text{t value} = \dfrac{145.9894}{2.71384} = 53.79$ et sa *p value* associée est bien inférieure au seuil 0.05.
On rejette l'hypothèse que $\beta_O = 0$ avec une grande confiance.

Même raisonnement pour l'estimateur de β_1 qui a pour valeur 0.17030.

Note
Dans le cas de la régression simple la statistique de test de l'estimateur de β_1 et lié à F : $F = (\text{t value})^2$

Dans la table en sortie par l'option outest=, SAS enregistre RMSE et les valeurs des paramètres.
SAS n'imprime pas en standard les intervalles de confiance des paramètres mais on peut les récupérer dans cette table en sortie, en utilisant l'option outest= et le mot clé Tableout.

Programme SAS

```
Proc REG data=libreg.tailpoid outest=TableSortie Tableout;
title 'Régression de la Taille en fonction du Poids ';
model y=x  ;
proc PRINT data=TableSortie;
title "Table produite par l'option outest avec le mot clé
Tableout";
run;
```

Table produite par l'option outest avec le mot clé Tableout

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	X	Y
1	MODEL1	PARMS	Y	2.57359	145.990	0.17030	-1
2	MODEL1	STDERR	Y	2.57359	2.714	0.02617	.
3	MODEL1	T	Y	2.57359	53.795	6.50803	.
4	MODEL1	PVALUE	Y	2.57359	0.000	0.00000	.
5	MODEL1	L95B	Y	2.57359	140.288	0.11532	.
6	MODEL1	U95B	Y	2.57359	151.691	0.22528	.

Les lignes L95B et U95B donnent les intervalles de confiance à 95% des paramètres.

1.4.4. Précision sur l'estimation de Y

On a vu que pour chaque valeur X_i fixée, la **vraie droite de régression** était le lieu de l'espérance (i.e. la **valeur moyenne)** de Y et que les Y devaient théoriquement se distribuer selon une loi normale centrée sur cette droite avec une variance théorique σ^2.

Pour évaluer la précision sur l'estimation de Y on aura deux optiques à considérer, soit on s'intéressera à l'intervalle de confiance autour de l'estimation de la droite de régression, soit on s'intéresse à l'intervalle de prévision de Y en fonction de X.

Intervalle de confiance autour de l'estimation de la droite de régression

Soit X_k représentant un niveau particulier de X pour lequel nous voulons estimer la valeur moyenne de Y. X_k peut être une valeur connue dans l'échantillon, ou une autre valeur de la variable régresseur non repérée dans l'échantillon. La réponse moyenne quand $X=X_k$ est notée $E(Y_k)$. L'estimateur de $E(Y_k)$ est noté $\hat{Y}_k$.

Il faut voir la distribution d'échantillonnage de $\hat{Y}_k$, comme la distribution que l'on obtiendrait si on effectuait des mesures répétées en X_k.

♦ **Calcul de l'erreur-type sur** $\hat{Y}_k$

On a vu que l'estimation de $E(Y_k)$ est donnée par :

$$\boxed{\hat{Y}_k = \overline{Y} + b_1(X - \overline{X}_k)}$$

Plaçons-nous en un point X_k et calculons la variance de $\hat{Y}_k$:

$$\boxed{Var(\hat{Y}_k) = Var(\overline{Y} + b_1(X_k - \overline{X}))}$$

Pour pouvoir calculer la variance il faut faire des suppositions sur les termes de cette expression.

Comme précédemment on suppose que les X_i sont non aléatoires.

Seuls la moyenne des Y_i et le coefficient b_1 sont des variables aléatoires. On peut montrer de plus que la covariance entre $\overline{Y}$ et le coefficient b_1 est nulle [11] .

Suppositions pour calculer la variance de $\hat{Y}_k$

Si les X_i sont non aléatoires

Si les Yi sont non corrélés et de même variance σ^2

Et comme par construction $Cov(\overline{Y}, b_1) = 0$

Alors :

$$\boxed{\begin{aligned} Var(\hat{Y}_k) &= Var(\overline{Y}) + (X_k - \overline{X})^2 Var(b_1) = \\ &= \frac{\sigma^2}{n} + (X_k - \overline{X})^2 \frac{\sigma^2}{\sum (X_i - \overline{X})^2} \end{aligned}}$$

Comme précédemment, on ne connaît pas la variance théorique σ^2 de Y. Il faut l'estimer.

[11] voir démonstration dans NETER, WASSERMAN, KUTNER pp75-77.

Supposition

Si le modèle postulé est le modèle correct
Alors σ^2 peut être estimé par les erreurs entre les Y observés et $\hat{Y}$

$$s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = MSE$$

L'estimateur de l'erreur-type de $\hat{Y}_k$ devient : $s(\hat{Y}_k) = s\left[\frac{1}{n} + \frac{(X_k - \overline{X})^2}{\sqrt{\sum (X_i - \overline{X})^2}}\right]^{1/2}$

♦ **Calcul de l'intervalle de confiance de** $\hat{Y}_k$

On montre que pour un modèle de régression la statistique $\frac{\hat{Y}_k - E(Y)}{s(\hat{Y}_k)}$ suit une distribution de Student à (n-2) degrés de liberté.
La **vraie valeur moyenne** μ_k de Y pour un X_k a une probabilité égale à (1-α) d'appartenir à l'intervalle de confiance :

$$IC_{1-\alpha}(E(Y_k)) = \left[\hat{Y}_k - t_{1-\alpha/2}.s(\hat{Y}_k); \hat{Y}_k + t_{1-\alpha/2}.s(\hat{Y}_k)\right]$$

L'intervalle de confiance de $\hat{Y}_k$ se matérialise par deux lignes courbes, des hyperboles, comme le montre la figure 1.9.

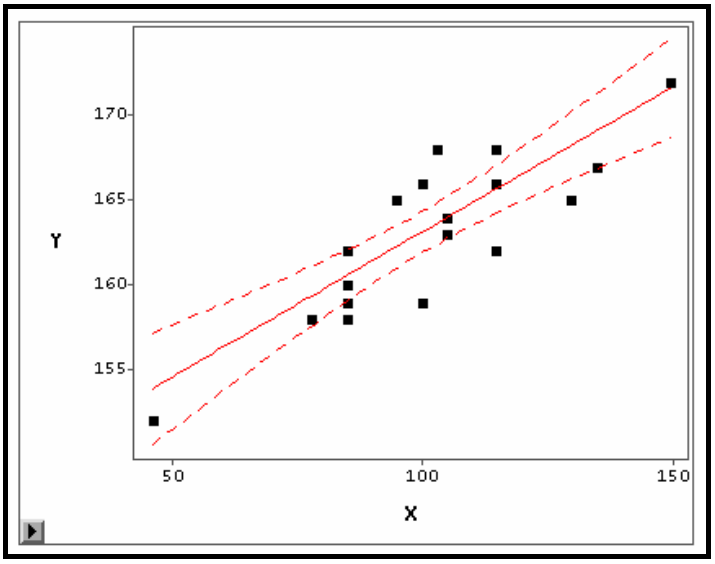


Figure 1.9 Intervalle de confiance à 95% de la moyenne des Tailles selon les valeurs des Poids

A propos de la largeur de l'intervalle de confiance, on peut faire les remarques suivantes :

- La largeur varie en fonction de $(X_k - \overline{X})$
- La largeur est minimum au point $X_k = \overline{X}$

  C'est dire que la précision est la **meilleure**, au **centre de gravité** du nuage des points
- La largeur croît lorsqu'on s'éloigne du centre de gravité. La précision est la plus **mauvaise** aux **extrémités** du nuage de points.

Intervalle de prévision de Y sachant X

Ici on s'intéresse à la prévision d'une *nouvelle* observation individuelle de Y pour une valeur X_k **,** de la variable X et non pas à la valeur moyenne de Y.

Dans ce cas, la variance de Y a deux composantes :

1. la variance de la position centrale de la distribution d'échantillonnage de $\hat{Y}_k$, cf. calcul réalisé au paragraphe précédent
2. la variance σ^2 de la distribution de Y autour de sa position centrale au point $X = X_k$. Comme précédemment, on estime σ^2 par s^2.

Pour une explication visuelle de cette décomposition[12] voir la figure 1.10.

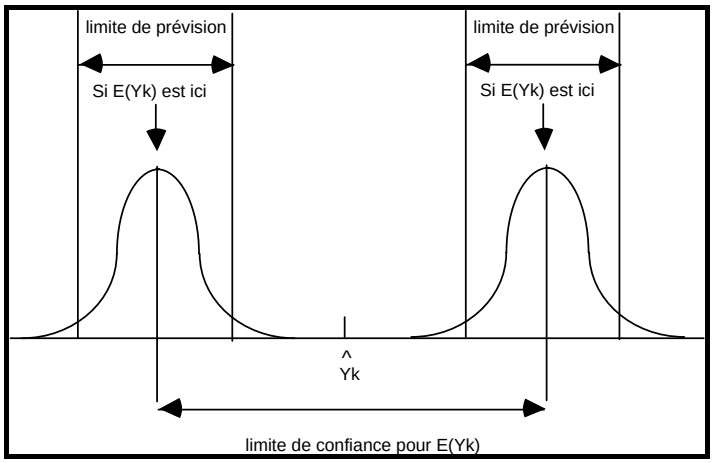


Figure 1.10 Illustration de la prédiction d'une nouvelle observation individuelle de Y

L'estimateur de l'erreur-type de Y sachant X devient :

$$s^2 + s^2(\hat{Y}_k) = s^2\left[1 + \frac{1}{n} + \frac{(X_k - \overline{X})^2}{\sqrt{\sum (X_i - \overline{X})^2}}\right]$$

L'intervalle de confiance d'une prévision de Y sachant X se matérialise là aussi par deux lignes courbes décalées d'une distance "s" par rapport à l'intervalle de confiance calculé pour la moyenne de Y_k.

[12]Source : NETER, WASSERMAN et KUTNER, p82.

Les remarques faites précédemment sur l'estimation de la moyenne de Y_k sont les mêmes que celles faites pour une observation individuelle. A savoir, la largeur de l'intervalle de confiance varie en fonction de $(X_k - \overline{X})$, c'est au centre de gravité du nuage de points que la précision est la meilleure, et aux extrémités du nuage de points que cette précision est la plus mauvaise.

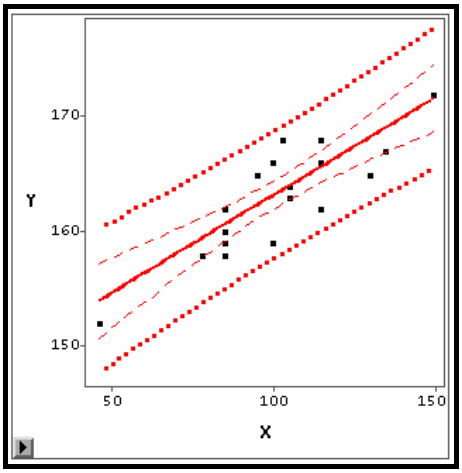


Figure 1.11 Intervalle de confiance à 95% des prévisions individuelles des Tailles

Sur la figure 1.11 on voit que l'intervalle de confiance des prévisions individuelles est évidemment plus grand que l'intervalle de confiance des moyennes théoriques.

Attention
En prévision et dans un cadre temporel, on cherche à estimer aux extrémités de la plage de variation de X, or c'est justement là que la précision est la moins bonne!

Exemple avec les options CLI CLM de la Proc REG

Les options CLI (Confidence Limit Individual) et CLM (Confidence Limit Mean) de l'instruction **model** de Proc REG donnent ces intervalles de confiance.
Pour sauvegarder ces valeurs dans une table SAS il faut utiliser l'instruction Output.

Programme SAS

```
Proc REG data=libreg.tailpoid ;
title 'Régression de la Taille en fonction du Poids ';
model y=x /CLI CLM  ;
Output Out=Table2 Predicted=Pred residual=Residu
       LCL=Borne_Inf_ind UCL=Borne_Sup_Ind
       LCLM=Borne_Inf_Moy UCLM=Borne_Sup_Moy;
proc PRINT data=Table2 ;title "Table produite par
l'instruction OUTPUT";
run;
```

Sortie de PROC REG

Régression de la Taille en fonction du Poids

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	152.0000	153.8238	1.5585	150.5495	157.0982	147.5027	160.1449	-1.8238
2	158.0000	159.2735	0.8394	157.5100	161.0369	153.5862	164.9607	-1.2735
3	160.0000	160.4656	0.7171	158.9590	161.9721	154.8527	166.0785	-0.4656
4	162.0000	160.4656	0.7171	158.9590	161.9721	154.8527	166.0785	1.5344
5	158.0000	160.4656	0.7171	158.9590	161.9721	154.8527	166.0785	-2.4656
6	159.0000	160.4656	0.7171	158.9590	161.9721	154.8527	166.0785	-1.4656
7	165.0000	162.1686	0.5990	160.9102	163.4270	156.6172	167.7200	2.8314
8	165.0000	162.1686	0.5990	160.9102	163.4270	156.6172	167.7200	2.8314
9	166.0000	163.0201	0.5766	161.8088	164.2314	157.4792	168.5610	2.9799
10	159.0000	163.0201	0.5766	161.8088	164.2314	157.4792	168.5610	-4.0201
11	166.0000	163.0201	0.5766	161.8088	164.2314	157.4792	168.5610	2.9799
12	168.0000	163.5310	0.5771	162.3186	164.7434	157.9898	169.0722	4.4690
13	163.0000	163.8716	0.5833	162.6460	165.0972	158.3275	169.4157	-0.8716
14	164.0000	163.8716	0.5833	162.6460	165.0972	158.3275	169.4157	0.1284
15	168.0000	165.5746	0.6773	164.1516	166.9976	159.9836	171.1656	2.4254
16	166.0000	165.5746	0.6773	164.1516	166.9976	159.9836	171.1656	0.4254
17	162.0000	165.5746	0.6773	164.1516	166.9976	159.9836	171.1656	-3.5746
18	165.0000	168.1291	0.9451	166.1435	170.1147	162.3692	173.8891	-3.1291
19	167.0000	168.9806	1.0519	166.7706	171.1907	163.1395	174.8218	-1.9806
20	172.0000	171.5352	1.3971	168.6000	174.4704	165.3829	177.6874	0.4648

Sum of Residuals	0
Sum of Squared Residuals	119.22082
Predicted Residual SS (PRESS)	142.46703

Lecture :

Les options CLM CLI donne pour chaque observation, les valeurs :

Dependant variable : Y

Predicted Value : $\hat{Y}$

Std Error mean predict : erreur-type au point X_i

95% CL Mean : les 2 colonnes suivantes donnent les bornes inférieure et supérieure de l'intervalle de prédiction à 95% de la moyenne.

95% CL Predict : les 2 colonnes suivantes donnent les bornes inférieure et supérieure de l'intervalle pour une prédiction individuelle.

Residual : résidu

L'instruction Output avec les mots clés LCL UCL LCLM UCLM permettent de récupérer ces statistiques dans une table SAS:

Table produite par l'instruction OUTPUT 3

Obs	I	X	Y	Pred	Borne_ Inf_Moy	Borne_ Sup_Moy	Borne_ Inf_ind	Borne_ Sup_Ind	Residu
1	1	46	152	153.824	150.549	157.098	147.503	160.145	-1.82381
2	2	78	158	159.273	157.510	161.037	153.586	164.961	-1.27346
3	3	85	160	160.466	158.959	161.972	154.853	166.078	-0.46557
4	4	85	162	160.466	158.959	161.972	154.853	166.078	1.53443
5	5	85	158	160.466	158.959	161.972	154.853	166.078	-2.46557
6	6	85	159	160.466	158.959	161.972	154.853	166.078	-1.46557
7	7	95	165	162.169	160.910	163.427	156.617	167.720	2.83141
8	8	95	165	162.169	160.910	163.427	156.617	167.720	2.83141
9	9	100	166	163.020	161.809	164.231	157.479	168.561	2.97991
10	10	100	159	163.020	161.809	164.231	157.479	168.561	-4.02009

2. La régression linéaire multiple

Dans ce chapitre nous reprenons les concepts de la régression linéaire simple pour les formaliser et les étendre à la régression multiple. Nous présentons les différentes formes de décomposition de sommes de carrés (Sum of Squares) et commentons les résultats obtenus avec la procédure REG.

2.1. Le critère des moindres carrés

Tout comme en régression linéaire simple; la régression linéaire multiple cherche à approximer une relation fonctionnelle trop complexe en général, par une fonction mathématique simple telle qu'une équation de la forme:

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon}$$

Reprenons le résumé des concepts de la régression linéaire présenté au chapitre 1.
L'équation de régression ou modèle postulé, met en relation:
• Y : variable **réponse** (à expliquer ou variable dépendante).
• Xj : variables **régresseurs** (explicatives ou variables indépendantes).

Cette équation est **linéaire** par rapport aux **paramètres** (coefficients de régression) $\beta_O, \beta_1, \cdots, \beta_p$. Le modèle est dit linéaire. Ces paramètres sont inconnus, on les estime en minimisant le critère des moindres carrés (**MCO** ou *Ordinary Least Squares*).
Le critère des moindres carrés correspond à la minimisation de la somme des carrés des écarts (SC Erreur en français, SS Error en anglais) entre *Y observé* et *Y estimé* par l'équation de régression.

Y estimé est noté $\hat{Y}$.

$$\boxed{\hat{Y}_i = b_0 + b_1 X_{1i} + \ldots + b_p X_{pi}}$$

avec:
Y : variable réponse
Xj : p variables régresseurs, j=1,…p
i indice de l'observation courante, i=1,…n
n le nombre d'observations.

Les valeurs qui minimisent ce critère sont des estimations $b_0, b_1, \ldots b_p$ des paramètres $\beta_O, \beta_1, \cdots, \beta_p$ inconnus.

Estimation des paramètres du modèle

Dans le cas d'un modèle à p variables régresseurs le critère des moindres carrés s'écrit:

$$S(\beta_0,\ldots\beta_p) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X1_i - \ldots\beta_p Xp_i)^2$$

Les valeurs des β qui minimisent ce critère seront les solutions b_0, b_1, …b_p du système linéaire de (p+1) équations à (p+1) inconnues.

$$\begin{array}{l} S_{11}b_1 + S_{12}b_2 + \ldots. + S_{1p}b_p = S_{1y} \\ \ldots. \\ S_{p1}b_1 + S_{p2}b_2 + \ldots. + S_{pp}b_p = S_{py} \end{array}$$

Avec

$$S_{kj} = \sum_{i=1,n}(X_{ki} - \overline{X}_k)(X_{ji} - \overline{X}_j) \quad \text{pour } k,j=1,2,\ldots p$$

$$S_{ky} = \sum_{i=1,n}(X_{ki} - \overline{X}_k)(Y_i - \overline{Y}) \quad \text{pour } k=1,2,\ldots,p$$

Pour résoudre un tel système linéaire les mathématiciens ont développé le calcul (algèbre) matriciel qui permet une présentation et des traitements compacts de grands tableaux de données. La notation matricielle est donc devenue l'unique moyen d'appréhender la régression multiple. Cependant cette présentation cache bien des difficultés du point de vue des résolutions numériques sur données réelles.

Les estimateurs des moindres carrés estiment les paramètres inconnus $\beta_O, \beta_1, \cdots, \beta_p$ avec une certaine précision. Sous les suppositions que les erreurs sont indépendantes et identiquement distribuées selon une loi normale, les estimateurs MCO sont centrés sur une valeur à laquelle est associé un intervalle de confiance. L'intervalle de confiance dépend de l'adéquation du modèle aux données, adéquation qui dépend des erreurs inconnues ε_i :

$$\varepsilon_i = Y_i - E(Y_i)$$

2.2. Formalisation de la régression linéaire multiple

En notation matricielle :

- **Y** est le vecteur colonne des n observations de la variable réponse
- **X**(n,p) la matrice des observations des p vecteurs X_i , chacun de dimension (n,1).

A cette matrice on ajoute en première colonne un vecteur constitué uniquement de 1. Ce vecteur correspond à la constante X_0. La matrice X est alors de dimension (n,p+1).
Cette représentation permet de traiter la constante X_0 comme une variable explicative.

- β est le vecteur colonne des (p+1) coefficients de régression ou paramètres inconnus β_i.
- ε représente le vecteur des erreurs.

$$Y = \begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \cdots \\ Y_n \end{vmatrix} \qquad X = \begin{vmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & & & & \\ 1 & & & & \\ \cdots & & & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{vmatrix} \qquad \beta = \begin{vmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{vmatrix} \qquad \varepsilon = \begin{vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdots \\ \varepsilon_n \end{vmatrix}$$

le modèle s'écrit: $\boxed{Y = X\beta + \varepsilon}$

Y estimé par le modèle de régression s'écrit: $\boxed{\hat{Y} = X\hat{\beta} = XB}$

Le vecteur colonne $\hat{\beta}$ (noté aussi B) représente le vecteur des estimateurs b_i des moindres carrés des paramètres inconnus β.
Les notations matricielles permettent d'écrire simplement le système à résoudre pour trouver les coefficients b_i qui minimisent le critère des moindres carrés:

$$\boxed{(X'X)B = (X'Y)}$$

X' désignant la matrice transposée de X.

Le vecteur B des coefficients solution s'obtient en inversant la matrice $(X'X)$:

$$\boxed{B = (X'X)^{-1}.(X'Y)}$$

La résolution de ce système n'est pas toujours possible. Cette résolution est liée à la possibilité d'inversion de la matrice $(X'X)$.

Supposons que 2 variables X_i et X_j soient corrélées entre elles c'est-à-dire qu'il existe une relation linéaire permettant de passer de X_i à X_j on a alors 2 lignes de la matrice $(X'X)$ qui sont proportionnelles et lorsque l'on veut résoudre le système il ne reste plus que p équations indépendantes et toujours (p+1) inconnues à trouver.
Le système est *indéterminé*, il existe une infinité de solutions.

Les variances des estimateurs (b) sont les éléments diagonaux de la matrice de variance-covariance des X inversée multipliés par la variance des erreurs σ^2 .

$$\boxed{\sigma^2(b) = \sigma^2(X'X)^{-1}}$$

Comme pour la régression simple σ^2 est estimé par $\boxed{MSE = \frac{SS\ error}{n-p-1}}$

Les variances des estimateurs dépendent des éléments diagonaux de la matrice à inverser. Si des régresseurs sont corrélés, les variances des estimateurs des paramètres sont élevées, et les estimations sont instables (non robustes). Un exemple de cette instabilité sera donné au chapitre 4.

La matrice H
A partir de l'expression du vecteur B des estimateurs des coefficients on peut calculer l'estimation de Y:

$$\hat{Y} = XB$$
$$\hat{Y} = X(X'X)^{-1}X'Y$$
$$\hat{Y} = HY$$
avec
$$H = X(X'X)^{-1}X'$$

Cette matrice H **-** *H comme Hat matrice***-** qui ne comporte que des données relatives aux variables régresseurs va jouer un rôle important, et son usage sera développé chapitre 4.

2.3. Exemples de régression linéaire multiple avec Proc REG

2.3.1. Présentation des données

Pour présenter la régression multiple avec quelques options de Proc REG, nous avons repris l'exemple de la chenille processionnaire du pin traité dans l'ouvrage de TOMASSONE & al. Cet exemple est fréquemment analysé dans la littérature française (voir FOUCART, AZAIS-BARBET). On pourra ainsi, avec leurs ouvrages, poursuivre des analyses plus complexes de ces données.
Le fichier de données est composé de 33 placettes où sont plantés des arbres infectés par des nids de chenille « procesionnaire du pin », une variable réponse (X11 et sa transformée en Log et dix variables régresseurs potentiels (X1-X10).
« Les expérimentateurs souhaitent connaître l'influence de certaines caractéristiques de peuplements forestiers (variables régresseurs X1-X10) sur le développement de la chenille processionnaire du pin (variable réponse X11 ou son logarithme) ».

X11 : Nombre de nids de processionnaires par arbre d'une placette. Log = Log(X11), transformation de la variable X11 par son logarithme X1 : Altitude (en mètre) X2 : pente (en degré) X3 : nombre de pins dans une placette de 5 ares X4 : hauteur de l'arbre échantillonné au centre de la placette X5 : diamètre de cet arbre

X6 : note de densité de peuplement
X7 : orientation de la placette
(1 orientation vers le sud, 2 autre)
X8 : Hauteur (en m) des arbres dominants
X9 : nombre de strates de végétation
X10 : mélange du peuplement (1 pas mélangé, 0 mélangé)

Données de base

Données Chenille processionnaire de TOMASSONE

Obs	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	log
1	1200	22	1	4.0	14.8	1.0	1.1	5.9	1.4	1.4	2.37	0.86289
2	1342	28	8	4.4	18.0	1.5	1.5	6.4	1.7	1.7	1.47	0.38526
3	1231	28	5	2.4	7.8	1.3	1.6	4.3	1.5	1.4	1.13	0.12222
4	1254	28	18	3.0	9.2	2.3	1.7	6.9	2.3	1.6	0.85	-0.16252
5	1357	32	7	3.7	10.7	1.4	1.7	6.6	1.8	1.3	0.24	-1.42712
6	1250	27	1	4.4	14.8	1.0	1.7	5.8	1.3	1.4	1.49	0.39878
7	1422	37	22	3.0	8.1	2.7	1.9	8.3	2.5	2.0	0.30	-1.20397
8	1309	46	7	5.7	19.6	1.5	1.3	7.8	1.8	1.6	0.07	-2.65926
9	1127	24	2	3.5	12.6	1.0	1.7	4.9	1.5	2.0	3.00	1.09861
10	1075	34	9	4.3	12.0	1.6	1.8	6.8	2.0	2.0	1.21	0.19062
11	1166	24	17	5.5	16.7	2.4	1.5	11.5	2.9	1.7	0.38	-0.96758
12	1182	41	32	5.4	21.6	3.3	1.4	11.3	2.8	2.0	0.70	-0.35667
13	1179	15	0	3.2	10.5	1.0	1.7	4.0	1.1	1.6	2.64	0.97078
14	1256	21	0	5.1	19.5	1.0	1.8	5.8	1.1	1.4	2.05	0.71784
15	1251	26	2	4.2	16.4	1.1	1.7	6.2	1.3	1.8	1.75	0.55962
16	1536	38	31	5.7	17.8	3.1	1.7	11.4	2.8	1.9	0.06	-2.81341
17	1554	27	20	5.6	20.2	2.8	1.9	9.2	2.7	1.3	0.13	-2.04022
18	1305	30	6	3.8	15.7	1.4	1.2	7.2	2.1	1.9	1.00	0.00000
19	1316	34	8	3.1	11.4	1.5	1.8	5.0	1.6	2.0	0.41	-0.89160
20	1427	39	19	4.6	15.2	2.4	1.6	9.1	2.4	1.9	0.72	-0.32850
21	1575	20	32	5.2	18.9	3.0	1.7	9.4	2.5	1.8	0.67	-0.40048
22	1397	26	16	4.2	14.8	2.2	1.6	7.7	2.2	1.8	0.12	-2.12026
23	1377	29	4	5.3	19.8	1.2	1.8	6.8	1.6	1.9	0.97	-0.03046
24	1574	24	23	5.2	17.8	2.4	1.8	7.8	2.2	2.0	0.07	-2.65926
25	1396	45	13	4.7	15.2	1.7	1.6	7.8	2.1	1.4	0.10	-2.30259
26	1393	27	5	4.7	18.3	1.2	1.7	7.5	1.7	2.0	0.68	-0.38566
27	1433	23	18	6.5	21.0	2.7	1.8	13.7	2.7	1.3	0.13	-2.04022
28	1349	24	1	2.7	5.8	1.0	1.7	3.6	1.3	1.8	0.20	-1.60944
29	1208	23	2	3.5	11.5	1.1	1.7	5.4	1.3	2.0	1.09	0.08618
30	1198	28	15	3.9	11.3	2.0	1.6	7.4	2.8	2.0	0.18	-1.71480
31	1228	31	6	5.4	21.8	1.3	1.7	7.0	1.5	1.9	0.35	-1.04982
32	1229	21	11	5.8	16.7	1.7	1.8	10.0	2.3	2.0	0.21	-1.56065
33	1310	36	17	5.2	17.8	2.3	1.9	10.3	2.6	2.0	0.03	-3.50656

2.3.2. Régression linéaire multiple avec Proc REG sans options

Nous étudions le modèle linéaire de la variable Log en fonction des 4 régresseurs X1, X2, X4, X5.

Etape 1 : Graphique de la matrice de diagrammes de dispersion (Scatter Plot avec SAS/INSIGHT)
Etape 2 : Analyse des corrélations entre les variables
Etape 3 : Régression multiple

Nous utilisons SAS/INSIGHT qui est beaucoup plus efficace pour obtenir des graphiques exploratoires, (voir en annexe 2 le mode d'emploi succinct de SAS/INSIGHT).

Programme SAS

```
/* étape 2 */
proc CORR data=libreg.chenilles;  title 'Corrélation de  X1 X2
X4 X5 avec Log';
var X1 X2 X4 X5 Log;
run;
/*étape 3 */
```

```
proc REG data=libreg.chenilles;
title 'Régression de LOG avec X1 X2 X4 X5 sans options';
model Log=X1 X2 X4 X5;
run;
```

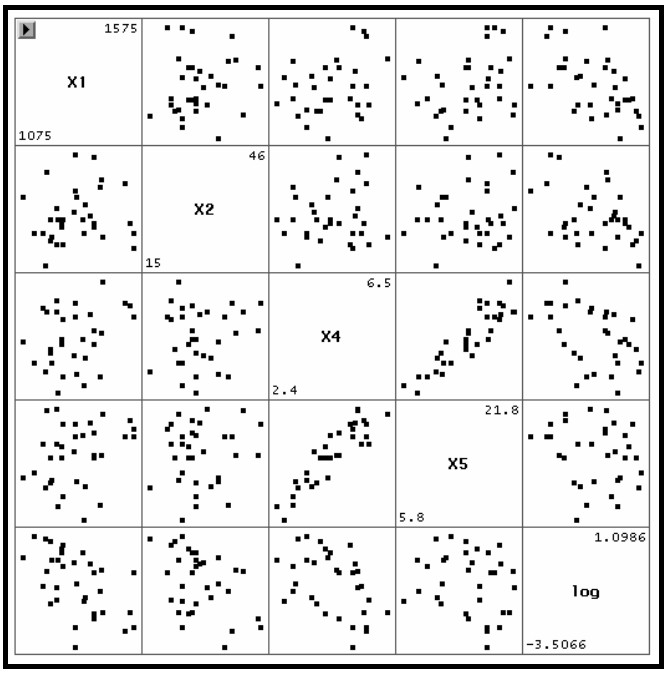


Figure 2.1: Matrice des diagrammes de dispersion des variables croisées 2*2.
Sur la diagonale sont affichées les valeurs min et max pour chaque variable.

Sortie SAS de PROC CORR

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X1	33	1315	129.03698	43406	1075	1575
X2	33	29.03030	7.30362	958.00000	15.00000	46.00000
X4	33	4.45152	1.04077	146.90000	2.40000	6.50000
X5	33	15.25152	4.30255	503.30000	5.80000	21.80000
log	33	-0.81328	1.24494	-26.83826	-3.50656	1.09861

Pearson Correlation Coefficients, N = 33
Prob > |r| under H0: Rho=0

	X1	X2	X4	X5	log
X1	1.00000	0.12052 0.5041	0.32105 0.0685	0.28377 0.1095	-0.53361 0.0014
X2	0.12052 0.5041	1.00000	0.13669 0.4481	0.11342 0.5297	-0.42944 0.0126
X4	0.32105 0.0685	0.13669 0.4481	1.00000	0.90466 <.0001	-0.42529 0.0136
X5	0.28377 0.1095	0.11342 0.5297	0.90466 <.0001	1.00000	-0.20094 0.2622
log	-0.53361 0.0014	-0.42944 0.0126	-0.42529 0.0136	-0.20094 0.2622	1.00000

Le graphique des diagrammes de dispersion de la figure 2.1, donne une image des liaisons entre toutes les variables X1, X2, X4, X5, Log. On voit d'un coup d'œil que les variables X4 et X5 sont très liées. Le coefficient de corrélation vaut 0.90466. D'autre part la variable réponse Log est liée négativement à tous les régresseurs.
La matrice de Scatter Plot est un complément utile à l'analyse de la matrice des coefficients de corrélation. Elle permet aussi de repérer les points atypiques (*outliers*) en X et en Y.

Sortie SAS de PROC REG sans options

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	32.09265	8.02316	12.83	<.0001
Error	28	17.50338	0.62512		
Corrected Total	32	49.59603			

Root MSE	0.79065	R-Square	0.6471
Dependent Mean	-0.81328	Adj R-Sq	0.5967
Coeff Var	-97.21683		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.73214	1.48858	5.19	<.0001
X1	1	-0.00392	0.00115	-3.42	0.0019
X2	1	-0.05734	0.01939	-2.96	0.0062
X4	1	-1.35614	0.31983	-4.24	0.0002
X5	1	0.28306	0.07626	3.71	0.0009

Cette sortie est analogue à celle de la régression simple, on retrouve les mêmes informations explicitées au chapitre1.

Lecture du test global dans le tableau de l'Analyse de Variance

- F value = 12.83 avec *p value* <0.001 ➔ rejet de H_0 tous les paramètres ne sont pas tous nuls

- R^2 = 0.6471 ➔ 64% de la variabilité de Y est expliquée par le modèle.

Lecture des paramètres

- Intercept b_0 = 7.73214 $s(b_0)$=1.48858
- Coefficient de X1 b_1=-0.00392 $s(b_1)$=0.00115
- Coefficient de X2 b_2=-0.05734 $s(b_2)$=0.01939
- Coefficient de X4 b_4= -1.35614 $s(b_4)$=0.31983
- Coefficient de X5 b_5= 0.28306 $s(b_5)$=0.07626

Toutes les *p-value* associées aux estimateurs des paramètres sont <0.05, on rejette l'hypothèse de nullité pour chacun des coefficients de la régression.
Cependant le coefficient de X5 est positif alors que la corrélation (X5, Log) est négative.
Peut-on alors parler d'un effet positif de cette variable X5 sur la variable réponse Log ?
La corrélation entre X4 et X5 provoque une instabilité des valeurs des coefficients.

2.4. TYPE I SS et TYPE II SS de Proc REG

Nous verrons d'abord la définition des statistiques Type I SS et Type II SS relatif à un paramètre puis les tests partiels relatifs à plusieurs paramètres.

2.4.1. Définition de TYPE I SS et TYPE II SS

Reprenons l'équation fondamentale de l'analyse de la variance :

SS Total = SS Model + SS Error

SS total qui représente la somme des carrés des écarts entre Y et sa moyenne **est invariant** quel que soit le nombre de variables régresseurs p dans le modèle. Lorsqu'on introduit une nouvelle variable régresseur dans un modèle SS Model augmente et donc SS Error diminue de la même quantité.
Pour juger de la contribution d'une variable régresseur à la réduction de SS Error, la Proc REG calcule pour chaque variable du modèle, deux sortes de SS Error.

- TYPE I SS : représente la réduction de SS Error liée à la variable lorsqu'elle est introduite **séquentiellement** dans le modèle.
- TYPE II SS : représente la réduction de SS Error liée à la variable lorsqu'elle est introduite la **dernière** dans le modèle.

• TYPE I SS
Soit le modèle complet définit avec les p régresseurs de l'instruction MODEL de Proc REG :

$$IC_{0.95}(\beta_i)=[b_i - 1.96 \cdot s(b_i); b_i + 1.96 \cdot s(b_i)]$$

Soit d'autre part le modèle restreint aux **<u>k premiers régresseurs</u> :**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Pour la "k"ème variable, TYPE I SS correspond à la différence entre SSerror du modèle à (k-1) régresseurs et SSerror du modèle à k régresseurs.

Attention : Le TYPE I SS d'une variable dépend de l'ordre de la variable dans l'instruction MODEL de Proc REG.

A TYPE I SS peut être associée une statistique de test, la *F Value* ou F de Fisher-Snedecor, et sa *p-value,* niveau de significativité du test. Le calcul de la *F value* et de sa *p value* associée n'est pas réalisé dans Proc REG[13].

• F VALUE : statistique de Fisher-Snedecor
Cette statistique de test vaut : $F\,VALUE = \frac{TYPE\,I\,SS}{MS\,Error}$

Le numérateur TYPE I SS correspond à la réduction de SS error lorsque l'on passe du modèle à (k-1) régresseurs -la variable étudiée étant exclue- au modèle à k régresseurs.
Le dénominateur MS Error correspond au modèle complet à p régresseurs .
La statistique de test *F value* ainsi définie permet de tester l'hypothèse nulle de la $k^{ième}$ variable.

[13] La statistique de Fisher-Snedecor *F value* et son niveau de significativité *p-value pour* Type I SS et TYPE II SS sont disponibles dans Proc GLM et dans SAS/INSIGHT.

Hypothèse à tester
On veut tester si le paramètre $\beta_k = 0$.

$H_0 : \beta_k = 0$ contre

$H_a : \beta_k \neq 0$

La statistique de test F value est sous H_0 une valeur observée d'une variable F de Fisher-Snedecor à 1 et (n-p-1) degrés de liberté. L'hypothèse nulle doit être rejetée au niveau α lorsque :

$$\boxed{\text{Fobservé} \geq F_{1-\alpha}(1, n-p-1)}$$

où $F_{1-\alpha}(1, n-p-1)$ représente le quantile d'ordre $(1-\alpha)$ de la loi de Fisher-Snedecor à (1) et (n-p-1) degrés de liberté.

Règle de décision

Si Fvalue $\geq F_{1-\alpha}(1, n-p-1)$
Alors H_0: $\beta_k = 0$ doit être rejeté au niveau α

Raisonnement
Au seuil α% la valeur maximum atteinte par F sous l'hypothèse nulle $H_0 : \beta_k = 0$ est $F_{1-\alpha}(1, n-p-1)$, si donc *F value* est supérieure on rejette l'hypothèse nulle. La variable contribue significativement à la réduction de SS Error, lorsqu'elle est entrée en **dernier** dans le modèle à k régresseurs.

- **Prob > F**

C'est la *p-value* associée à *F value.*
On compare la *p-value*, au risque α choisi (par exemple α=0.05).

Raisonnement sur la p-value

Si p-value $\leq \alpha$
Alors on rejette l'hypothèse nulle $\beta_k = 0$

Interprétation : Si la probabilité (Prob>F) est faible (<0.05) la variable contribue significativement à la réduction de SS Error dans le modèle à k régresseurs.

- **TYPE II SS**

Soit le modèle complet définit avec p régresseurs dans l'instruction MODEL de Proc REG :

$$\boxed{IC_{0.95}(\beta_i) = [b_i - 1.96 \cdot s(b_i); b_i + 1.96 \cdot s(b_i)]}$$

Pour la $k^{ième}$ variable, TYPE II SS correspond à la différence entre SSerror du modèle à **(p-1) régresseurs** (le $k^{ième}$ régresseur étant exclu) et SSerror du modèle complet à p régresseurs.

A TYPE II SS peut être associée une statistique de test, la *F value* ou F de Fisher-Snedecor, et sa *p-value,* niveau de significativité du test.

Remarque : Par construction TYPE I SS et TYPE II SS de la dernière variable du modèle ont la même valeur.

Le calcul de la *F value* et de sa *p value* associée n'est pas réalisé dans Proc REG, il faut faire le calcul à la main, ou utiliser l'option de Proc REG, SELECTION= FORWARD ou BACKWARD qui donne les *F value* de TYPE II à chaque pas .

• F VALUE : statistique de Fisher-Snedecor

Cette statistique de test vaut : $\boxed{\text{F VALUE} = \frac{\text{TYPE II SS}}{\text{MS Error}}}$

Le numérateur TYPE II SS correspond à la réduction de SS error lorsque l'on passe du modèle à (p-1) régresseurs –le régresseur étudié étant exclu- au modèle complet à p régresseurs.
Le dénominateur MS Error correspond au **modèle complet**, avec les p régresseurs .
La statistique de test *F value* ainsi définie permet de tester l'hypothèse nulle du "k"ème régresseur, lorsqu'il entre en dernier dans le modèle. C'est un F dit **partiel**.

Hypothèse à tester

On veut tester si le paramètre $\beta_k = 0$.

$H_0 : \beta_k = 0$ contre

$H_a : \beta_k \neq 0$

La statistique de test F value est sous H_0 une valeur observée d'une variable F de Fisher-Snedecor à 1 et (n-p-1) degrés de liberté. L'hypothèse nulle doit être rejetée au niveau α lorsque :

$$\boxed{\text{Fobservé} \geq F_{1-\alpha}(1, n-p-1)}$$

où $F_{1-\alpha}(1, n-p-1)$ représente le quantile d'ordre $(1-\alpha)$ de la loi de Fisher-Snedecor à (1) et (n-p-1) degrés de liberté.

Règle de décision

Si Fvalue $\geq F_{1-\alpha}(1, n-p-1)$
Alors $H_0 : \beta_k = 0$ doit être rejetée au niveau α

Raisonnement
Au seuil α% la valeur maximum atteinte par F sous l'hypothèse nulle H_0 : $\beta_k = 0$ est $F_{1-\alpha}(1, n-p-1)$, si donc *F value* est supérieure on rejette l'hypothèse nulle.
Le régresseur contribue significativement à la réduction de SS Error, lorsqu'il est entré en **dernier** dans le modèle.

Remarque : *F value* = T^2, avec T représentant la valeur du test de Student associé au paramètre.

- **Prob > F**

C'est la *p-value* associée à *F value.* On compare la *p-value*, au risque α choisi (par ex : α=0.05).
Raisonnement sur la p-value

Si p-value $\leq \alpha$ Alors on rejette l'hypothèse nulle $\beta_k = 0$

Interprétation : Si la probabilité (Prob>F) est faible, le régresseur contribue significativement à la réduction de SS Error, même lorsqu'il est entré en **dernier** dans le modèle complet à p régressseurs.

2.4.2. Interprétations conjointes de TYPE I SS et TYPE II SS

- Lorsque le modèle ne comporte qu'une seule variable régresseur :
  TYPE I SS = TYPE II SS = SS modèle

- Lorsque le modèle comporte plus d'une variable régresseur :

le TYPE I SS (lié au F séquentiel) dépend de l'ordre d'apparition des variables régresseurs dans l'instruction MODEL, tandis que TYPE II SS (lié au F partiel) n'en dépend pas.

Si pour un régresseur X_i, le F séquentiel et le F partiel sont plus grands que ceux des autres régresseurs, alors X_i a une **contribution plus grande** puisqu'il réduit plus la variation de SS Error, que la variable soit entrée en séquence dans le modèle ou en dernier.

Si pour un régresseur X_i, le F séquentiel est significatif et le F partiel ne l'est plus c'est qu'il y a des **colinéarités entre les régresseurs.**

L'information apportée par ce régresseur est redondante par rapport à l'information apportée par les précédents régresseurs déjà introduits dans le modèle.

2.4.3. Options SS1 et SS2 de l'instruction model de Proc REG

Les options SS1 et SS2 de l'instruction model de PROC REG permettent d'obtenir les statistiques Type I SS et Type II SS.

Programme SAS

```
proc REG data=libreg.chenilles;
     title 'Régression de Log avec X1 X2 X4 X5 avec Options
SS1 SS2 ';
     model Log=X1 X2 X4 X5/ SS1 SS2;
run;
```

Sortie de Proc REG de SAS

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	7.73214	1.48858	5.19	<.0001	21.82704	16.86620
X1	1	-0.00392	0.00115	-3.42	0.0019	14.12216	7.30671
X2	1	-0.05734	0.01939	-2.96	0.0062	6.70953	5.46832
X4	1	-1.35614	0.31983	-4.24	0.0002	2.64867	11.23888
X5	1	0.28306	0.07626	3.71	0.0009	8.61230	8.61230

Lecture

Exemple pour X1 :

- Type I SS = 14.12216 est la réduction de SS error lorsque la variable X1 est entrée la première dans le modèle (elle est alors la seule variable régresseur).
- Type II SS =7.30671 est la réduction de SS error lorsque la variable X1 est entrée la dernière dans le modèle.

Pour tester si cette réduction est significative il faut faire le calcul à la main, car dans cette sortie, Proc REG ne fournit pas les F value et les proba associées pour TYPE I SS et TYPE II SS :

$$\text{F value} = \frac{\text{TYPE I SS}}{\text{MS Error}} = \frac{14.1222}{0.62512} = 22.59$$

$$\text{F value} = \frac{\text{TYPE II SS}}{\text{MS Error}} = \frac{7.30671}{0.62512} = 11.69$$

On peut vérifier ces F value avec les sorties de SAS/INSIGHT.

Sortie avec SAS Insight

Type I Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	14.1222	14.1222	22.59	<.0001
X2	1	6.7095	6.7095	10.73	0.0028
X4	1	2.6487	2.6487	4.24	0.0490
X5	1	8.6123	8.6123	13.78	0.0009

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	7.3067	7.3067	11.69	0.0019
X2	1	5.4683	5.4683	8.75	0.0062
X4	1	11.2389	11.2389	17.98	0.0002
X5	1	8.6123	8.6123	13.78	0.0009

Note : SAS/ INSIGHT nomme TYPE III tests ce qu'on a appelé TYPE II SS dans la Proc REG

On retrouve bien les calculs faits à la main (cf. **F Stat =22.59** pour la F value de Type I SS et **F Stat=11.69** pour la F value de Type II SS).

La variable X4 a un comportement bizarre, lorsqu'elle est entrée en 3ème rang dans le modèle elle est limite au niveau significativité (pvalue =0.0490), alors que son apport est **très significatif** lorsqu'elle est entrée la dernière (p value =0.0002)

La liaison entre X4 et X5 nous joue des tours !

2.4.4. Tester la nullité de r paramètres pour tester un sous modèle

Ce type d'analyse est d'usage courant en Econométrie. L'idée est de mettre à l'épreuve une approche **théorique** par une validation **empirique**. L'intérêt porte non sur l'estimation des paramètres mais sur la « spécification » du modèle. Par spécification on entend la recherche des variables-régresseurs intervenant dans la détermination de la variable « à expliquer » Y.

On veut tester la nullité de r (indices k à q) paramètres parmi les p, c'est à dire l'hypothèse nulle

H_0 : $\beta_k = \cdots = \beta_q = 0$ contre

H_a : il y a parmi les $\beta_k \cdots \beta_q$ des coefficients non égaux à 0) *(not all* $\beta_k \cdots \beta_q$ *equal to 0)*.

Le modèle sans les r variables est appelé le **modèle restreint** par opposition au **modèle complet** à p variables.

Ici aussi on raisonne sur les réductions de SS error.
On note :
RRSS (Restricted Residual Sum of Squares) = Somme des carrés des résidus du modèle restreint
URSS (Unrestricted Residual Sum of Squares)=Somme des carrés des résidus du modèle complet.

L'hypothèse est testée en évaluant la statistique F dite partielle[14]

$$F = \frac{(RRSS - URSS)/r}{URSS/(n-p-1)} = \frac{(RRSS - URSS)/r}{MSE}$$

La statistique de test F est sous H_0 une valeur observée d'une variable F de Fisher-Snedecor à r et (n-p-1) degrés de liberté. L'hypothèse nulle doit être rejetée au niveau α lorsque :

$$\text{Fobservé} \geq F_{1-\alpha}(r, n-p-1)$$

où $F_{1-\alpha}(r, n-p-1)$ représente le quantile d'ordre $(1-\alpha)$ de la loi de Fisher-Snedecor à (r) et (n-p-1) degrés de liberté.

Règle de décision

Si $\text{Fvalue} \geq F_{1-\alpha}(r, n-p-1)$
Alors $H_0 : \beta_k = \cdots = \beta_q = 0$ doit être rejetée au niveau α

L'instruction **TEST** de la Proc REG réalise ces tests en fournissant la statistique de Fisher-Snedecor *F value* et son niveau de significativité *p-value* associé, noté *Prob>F*.

2.4.5. Exemple de test partiel avec PROC REG

On veut tester si les 2 coefficients de X4 et X5 sont nuls.

$H_0 : \beta_4 = \beta_5 = 0$ instruction SAS ⇨ `test X4=0, X5=0;`

Programme SAS

```
proc REG data=libreg.chenilles;
      title "Test de l'Hypothèse nulle X4=0 et X5=0";
      model Log=X1 X2 X4 X5;
test X4=0, X5=0;
```

- Pour le modèle restreint (sans X4 et X5) on a :

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	20.83168	10.41584	10.86	0.0003
Error	30	28.76434	0.95881		
Corrected Total	32	49.59603			

[14] La statistique de test F ainsi calculée est appelée *F partiel* quand elle ne porte que sur un sous-ensemble de paramètres, pour la distinguer de la statistique F, qui porte sur l'ensemble des paramètres du modèle complet.

➔ $\boxed{RRSS = 28.76434}$

- Pour le modèle complet :

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	32.09265	8.02316	12.83	<.0001
Error	28	17.50338	0.62512		
Corrected Total	32	49.59603			

➔ $\boxed{URSS = 17.50338}$ avec DF=(n-p-1)=28

d'où

$$\boxed{F = \frac{(RRSS - URSS)/r}{URSS/(n-p-1)} = \frac{(RRSS - URSS)/r}{MSE} = \frac{(28.76434 - 17.50338)/2}{17.80338/28} = \frac{5.63048}{0.62512} = 9.01}$$

Sortie SAS pour le test partiel

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable log

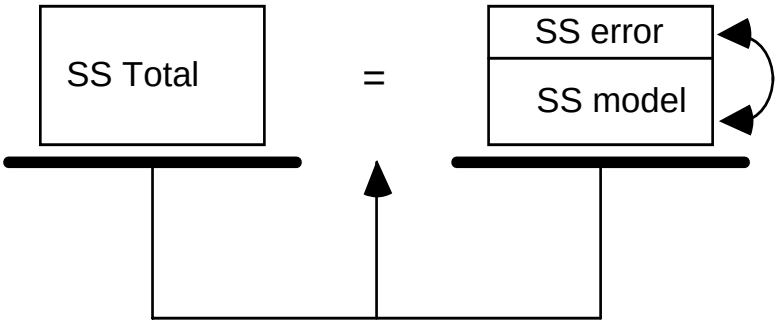
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	5.63048	9.01	0.0010
Denominator	28	0.62512		

On trouve bien la valeur $\boxed{F\ Value = 9.01}$ avec un niveau de significativité Prob >F de 0.0010.

Conclusion
Le niveau de significativité (0.0010) étant bien inférieur à 0.05, on rejette l'hypothèse nulle $\boxed{H_0 : \beta_4 = \beta_5 = 0}$
Il existe au moins un effet de X_4 et/ou de X_5 sachant X_1 et X_2 introduit dans le modèle.

2.5. Ce qu'il faut retenir des 'SS'

Décomposition des SS : Sum of Squares



Lorsqu'on introduit une nouvelle variable dans un modèle :

SS model augmente de la même quantité que

SS error décroît

Donc le coefficient de détermination R-square augmente toujours.

Cependant on n'améliore pas nécessairement la précision de l'estimation de Y.

En effet SS Error décroît mais $s^2 = \text{MSError} = \frac{\text{SS Error}}{n-p-1}$ peut croître donc augmenter la largeur de l'intervalle de confiance de Y estimé qui est proportionnel à MSE.

A la limite si le nombre de variables p + 1 (le 1 correspond à la variable constante X_0) est égal au nombre d'observations (n), l'équation de régression passera exactement par tous les points du nuage, l'ajustement sera parfait.

Dans ce cas SS Error vaut 0, et le coefficient de détermination R^2 vaut 1. Ce n'est plus de la statistique mais de la résolution d'équations !

D'autre part les colinéarités entre les régresseurs rendent les résultats instables. En augmentant le nombre de régresseurs on augmente les risques de colinéarités. Le chapitre 4 traitera de ce problème.

Modèles "parcimonieux"
Par sagesse, les statisticiens parlent de modèles « parcimonieux », pour signifier qu'un modèle doit comporter un nombre limité de variables par rapport au nombre d'observations, si on veut que le modèle ait une portée prévisionnelle et/ou explicative.

2.6. Les résidus

En un point d'observation i, l'écart entre Y observé et Y estimé par le modèle est le résidu au point i:

$$\boxed{e_i = Y_i - \hat{Y}_i}$$

Ces résidus e_i sont vus comme les erreurs observées des vraies erreurs inconnues ε_i :

$$\boxed{\varepsilon_i = Y_i - E(Y_i)}$$

Nous avons vu que les suppositions faites sur les ε_i pour élaborer les tests statistiques se résument ainsi « *les erreurs doivent être indépendantes et identiquement distribuées selon une loi normale* ».

Si le modèle est approprié aux données, les résidus observés e_i doivent refléter les propriétés des vraies erreurs inconnues ε_i.
C'est donc par le biais de l'analyse des résidus que l'on cherchera à valider le modèle de régression postulé.

Pour cela on effectuera différents graphiques des résidus en fonction de:

- Y la variable réponse
- Y estimé ($\hat{Y}$)
- X_i les variables régresseurs
- la variable temporelle, si l'analyse statistique porte sur des séries chronologiques
- etc.

De même on étudiera la normalité des résidus, leur indépendance.
En effet, les résidus contiennent d'une part un aléa d'espérance nulle et de variance σ^2, et d'autre part une information concernant l'inadéquation du modèle aux données (c'est-à-dire l'écart entre le modèle postulé et le modèle correct inconnu). Ce que l'on veut c'est que l'importance de cette deuxième partie soit moindre que celle due à l'aléa.

Pour cela on devra rechercher si dans les résidus il n'existe pas une structure organisée ou un contenu informationnel qui prouverait que le modèle postulé se différencie significativement du modèle correct.

Tous les tests sont faits en supposant que le modèle postulé est le modèle correct, si donc l'analyse des résidus prouve l'inadéquation du modèle postulé, les tests ne sont plus valables, ou sont biaisés.

Des orientations pour l'analyse critique des résidus seront données dans le chapitre 4.

Conclusion

Au cours des chapitres 1 et 2 nous avons présenté la majeure partie des concepts théoriques nécessaires à la compréhension d'un modèle de régression. Les démonstrations ont été limitées à l'essentiel. Le lecteur se reportera à la bibliographie pour avoir plus de précision et de rigueur mathématique. L'ouvrage de TOMASSONE & al., en particulier, est vivement conseillé.

Dans le chapitre 3, nous analyserons à partir d'exemples de la littérature, les difficultés de la régression linéaire lorsqu'on étudie des données réelles, qui « *prennent un malin plaisir* » à ne pas se comporter comme la théorie le suppose.

Pour aider aux diagnostics, de nombreuses options sont disponibles dans la procédure REG, ce que nous verrons au chapitre 4.

3. Quand les résultats d'une régression ne sont pas forcément pertinents

Dans ce chapitre nous montrons sur quelques exemples les difficultés rencontrées dans l'application de la régression linéaire simple et la régression linéaire multiple sur données réelles ou simulées, lorsque les suppositions ne sont pas vérifiées. Nous présentons quelques aides très utiles. Ce n'est qu'un petit survol de la littérature sur le sujet de la robustesse qui nécessite à lui seul plusieurs ouvrages.

Ce chapitre a pour objectif de vous sensibiliser à l'importance des diagnostics proposés dans Proc REG, qui seront vus au chapitre 4.
La majorité des résultats d'une régression sont présentés avec les sorties de SAS/INSIGHT, pour montrer l'apport de l'interactivité, et l'importance des graphiques dans la compréhension des analyses.

3.1. Exemples en régression simple

3.1.1. Une même valeur pour des situations différentes

Cet exemple est deTOMASSONE & *al.* (1986). Dès 1973 ANSCOMBE, neveu et collaborateur de J.W TUKEY (1977) avait proposé un exemple similaire dans «Graphs in Statistical Analysis ». Soient les 5 couples de 16 observations (X,Ya),(X,Yb),(X,Yc),(X,Yd),(Xe,Ye) sur lesquels on effectue 5 régressions linéaires simples.

OBS	X	Ya	Yb	Yc	Yd	Xe	Ye
1	7	5,535	0,113	7,399	3,864	13,715	5,654
2	8	9,942	3,770	8,546	4,942	13,715	7,072
3	9	4,249	7,426	8,468	7,504	13,715	8,491
4	10	8,656	8,792	9,616	8,581	13,715	9,909
5	12	10,737	12,688	10,685	12,221	13,715	9,909
6	13	15,144	12,889	10,607	8,842	13,715	9,909
7	14	13,939	14,253	10,529	9,919	13,715	11,327
8	14	9,450	16,545	11,754	15,860	13,715	11,327
9	15	7,124	15,620	11,676	13,967	13,715	12,746
10	17	13,693	17,206	12,745	19,092	13,715	12,746
11	18	18,100	16,281	13,893	17,198	13,715	12,746
12	19	11,285	17,647	12,590	12,334	13,715	14,164
13	19	21,365	14,211	15,040	19,761	13,715	15,582
14	20	15,692	15,577	13,737	16,382	13,715	15,582
15	21	18,977	14,652	14,884	18,945	13,715	17,001
16	23	17,690	13,947	29,431	12,187	33,281	27,435

Les estimations des Y sont les suivantes :

OBS	Ya_est	Yb_est	Yc_est	Yd_est	Ye_est
1	6,18	6,18	6,18	6,18	11,61
2	6,99	6,99	6,99	6,99	11,61
3	7,80	7,80	7,80	7,80	11,61
4	8,61	8,61	8,61	8,61	11,61
5	10,22	10,23	10,22	10,22	11,61
6	11,03	11,03	11,03	11,03	11,61
7	11,84	11,84	11,84	11,84	11,61
8	11,84	11,84	11,84	11,84	11,61
9	12,65	12,65	12,65	12,65	11,61
10	14,27	14,27	14,27	14,27	11,61
11	15,07	15,08	15,08	15,08	11,61
12	15,88	15,89	15,89	15,89	11,61
13	15,88	15,89	15,89	15,89	11,61
14	16,69	16,69	16,69	16,69	11,61
15	17,50	17,50	17,50	17,50	11,61
16	19,12	19,12	19,12	19,12	27,43

Les ajustements par les 5 droites de régressions sont donnés figure 3.1 ; ils sont identiques, mêmes estimations, mêmes statistiques pour R^2=0.617, mêmes coefficients b_0=0.520, b_1=0.809, mêmes erreurs-types sur les coefficients, et pourtant les situations sont bien différentes.

Analyse des résultats

- Sur le 1er graphique on peut voir que le modèle semble bien **adapté**.
- Sur le 2ième graphique le modèle linéaire est **inadapté**, un modèle quadratique de la forme $Y = \beta X^2$ serait préférable.
- Sur le 3ième graphique un point est **suspect** et entraîne la droite de régression vers le haut.
- Sur le 4ième graphique la variance des erreurs varie. Il y a un phénomène **d'hétéroscédasticité** (variance de Y sachant X non constante).
- Sur le 5ième graphique le **plan expérimental** défini par les valeurs de Xe est particulièrement mauvais. Un seul point extrême détermine la droite. Que se passe-t-il entre X=13.715 et X=34.281 ? La liaison est-elle linéaire ?

Un simple diagramme cartésien (Scatter plot) permet dans chacun des 5 cas de vérifier si les suppositions de la régression linéaire sont respectées, et de porter un diagnostic.

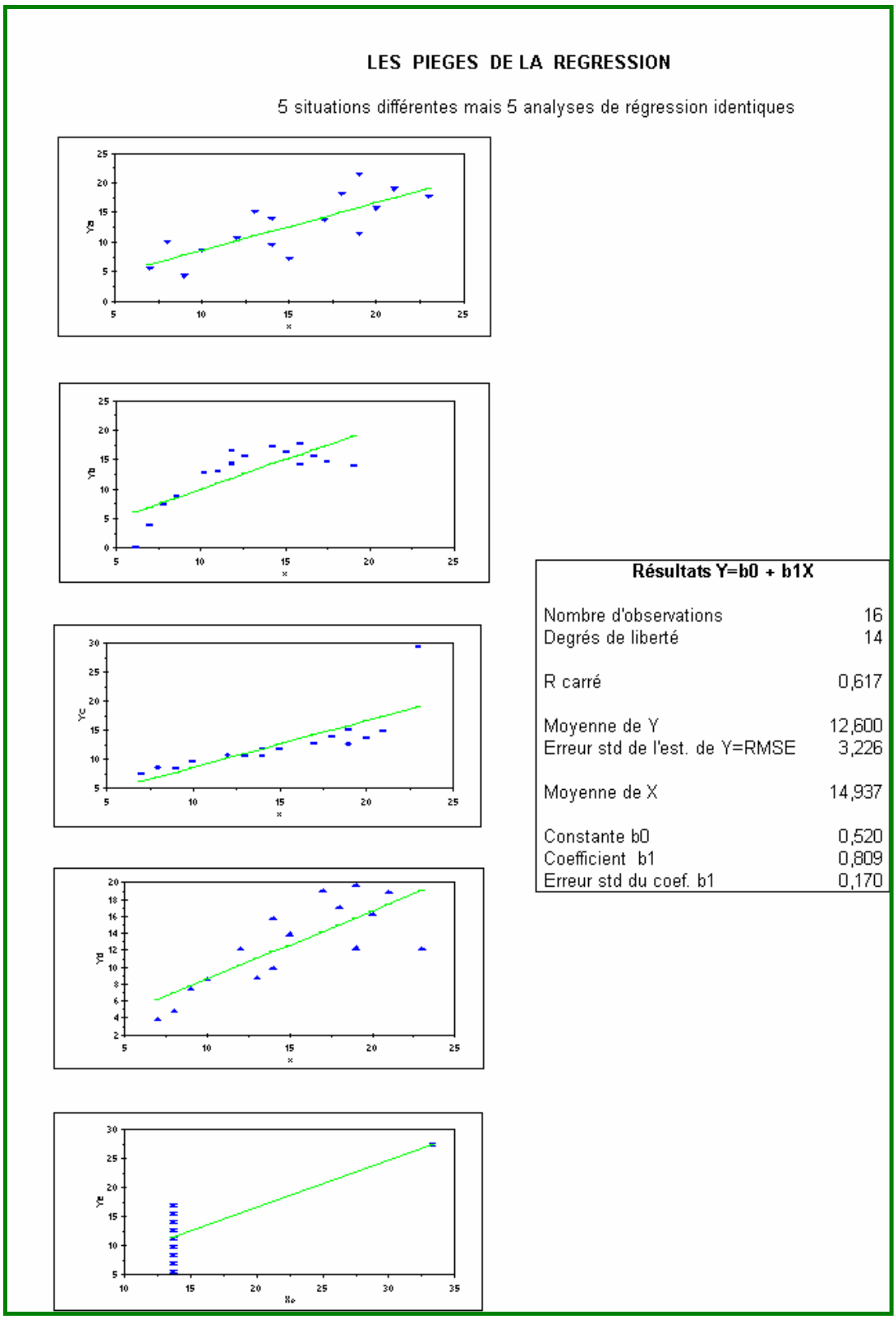


Figure 3.1 : 5 droites de régression

3.1.2. Pondérations et régression linéaire par morceaux

Cet exemple inspiré de J.P. BENZECRI & F.BENZECRI (1989), « Calculs de corrélation entre variables et juxtaposition de tableaux », est totalement artificiel.

L'objectif de cet exemple est double : montrer d'une part l'effet d'une pondération des observations sur les résultats d'une analyse de régression et d'autre part de sensibiliser par des graphiques, à l'usage abusif de la régression lorsque l'hypothèse de linéarité sur tout l'intervalle n'est pas valide.

Tableau des données
101 observations (variables X et Y) sont générées par programme de la manière suivante : X varie de -50 à +50 avec un pas de 1 et à chaque pas Y est calculé selon les formules linéaires suivantes :

$$\boxed{\begin{array}{l} \text{Si } X < -10 \text{ alors } Y = X + 20 \\ \text{Si } -10 \leq X < 11 \text{ alors } Y = -X \\ \text{Si } X \geq 11 \text{ alors } Y = X - 20 \end{array}}$$

X est la variable régresseur, Y est la variable réponse. X et Y sont donc *rigoureusement* liées par une fonction linéaire par morceaux.
Les 3 parties ont des pentes (paramètre β_1) respectives de +1 pour les 40 premières observations, -1 pour les observations 41 à 61, et +1 pour les observations 62 à 101.

On effectue une première régression sans pondération, sur toutes les observations (figure 3.2), puis deux régressions pondérées, en pondérant par 100 les observations de la **partie centrale,** les autres observations ayant une pondération de 1, puis une 3ième régression avec une pondération par 1000.
On utilise l'instruction "**Weight** "de la Proc Reg.

Programme SAS de génération des observations et analyse de régression
L'appel de Proc REG se fait par une macro.

```
data ligne ;
    do x=-50 to 50;
    p1=1 ;  p100=1; pmil=1;
    if x <-10 then  y=x+20;
              else if x < 11 then do; y=-x ;
                                  p100=100; pmil =1000; end;
                             else y=x-20;
    output;
end;
%macro reg(poids=);
proc reg data=ligne;
     model y=x ;
     %if &poids ^=  %then weight &poids ; %str(;);
     title " Avec ponderation &poids";
%mend;
%reg(poids=p1)
%reg(poids=p100)
%reg(poids=p1000)
quit;
```

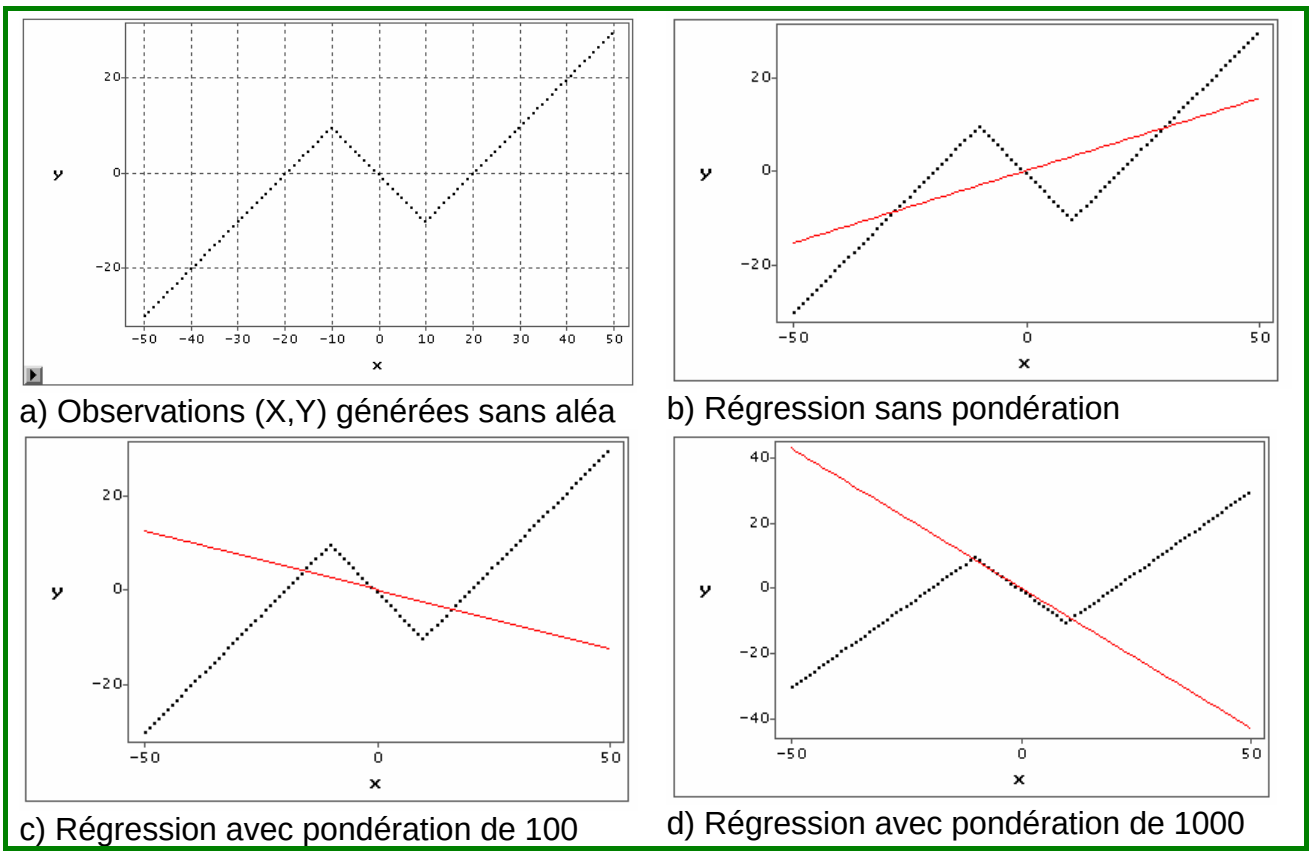


Figure 3.2 Droite de régression selon les pondérations utilisées (1, 100 , 1000)

Analyse des résultats

En pondérant de 3 façons différentes la partie centrale on obtient des *F value*, des coefficients de détermination, des estimations de β_1 droite de régression- totalement différents et cependant toujours significatif (cf. Tableau 3.1)

Tableau 3.1

Pondération	F value	P value	R^2	estimation de β_1	T Student
sans pondération	261.433	<0.0001	0.7253	0.0255	16.169
par 100	11.740	<0.0009	0.1060	-0.2512	-3.427
par 1000	389.810	<0.0001	0.7985	-0.8580	-19.744

En fonction de la pondération, la droite de régression « tourne » jusqu'à s'adapter à la pente de la partie centrale, comme on le voit dans les figures 3.2 (b, c, d).
Si la modélisation devient correcte pour la partie centrale, il n'en va pas de même pour les deux autres parties.

Si, par contre on réalise 3 études séparées sur les 3 intervalles où l'hypothèse de linéarité entre X et Y est exacte, on obtient des coefficients de détermination $R^2=1$ (corrélation parfaite) et des droites de régression totalement adaptées aux données (SS Error=0).

L'exemple met en lumière les questions que l'on doit se poser avant d'effectuer une régression : **l'hypothèse de linéarité est elle plausible sur tout l'intervalle ?**.

Théorie de la régression pondérée

● *L'instruction **Weight** de Proc REG*

L'instruction **Weight** de Proc REG, minimise la somme des carrés résiduels pondérés :

$$\sum_i w_i (Y_i - \hat{Y}_i)^2$$

w_i : valeur de la variable spécifiée dans l'instruction Weight,

Y_i : valeur observée de la variable à expliquer,

$\hat{Y}_i$: valeur prédite pour l'observation i.

Les équations normales utilisées sont dans ce cas :

$$\beta = (X'WX)^{-1}(X'WY)$$

W : matrice diagonale constituée des poids.

● *Quand utiliser l'instruction Weight ?*

Lorsque l'hypothèse de variance constante des erreurs n'est pas vérifiée (hétéroscédasticité des erreurs), la littérature statistique propose d'utiliser la régression pondérée en prenant comme pondération l'inverse des variances théoriques σ_i^2 .

$$w_i = \frac{1}{\sigma_i^2}$$

En général les variances théoriques des erreurs ne sont pas connues, elles sont estimées.

Conclusion

L'artifice de pondération des données doit être utilisé avec discernement. En particulier il faut au préalable contrôler si les observations sont homogènes, c'est-à-dire si d'éventuels groupes aux comportements différents peuvent être différenciés. Une approche par analyse de données ou analyse exploratoire des données peut s'avérer nécessaire. On utilisera avec profit des marqueurs de couleurs (SAS/INSIGHT et sa boîte à outils **Tools**) pour repérer si une variable de groupe ne fait pas apparaître des mélanges de populations.

Il existe des **méthodes robustes** pour pondérer les observations mais elles sortent du cadre de cet ouvrage. La lecture de l'ouvrage de ROUSSEEUW et al. (2003) est fortement recommandée.

3.1.3. Transformation des données

Lorsque sur un graphique, la liaison entre X et Y n'apparaît pas linéaire, on peut essayer de transformer les données, pour tenter de linéariser la liaison.

Exemple : Liaison entre Produit National Brut et taux d'urbanisation
La table SAS Paysniv3 porte sur 173 pays des 5 continents (figure 3.3).

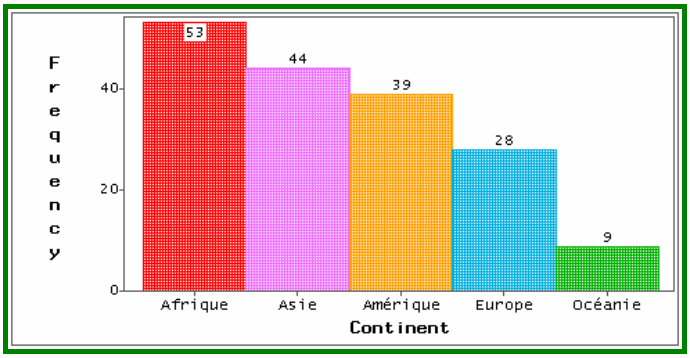
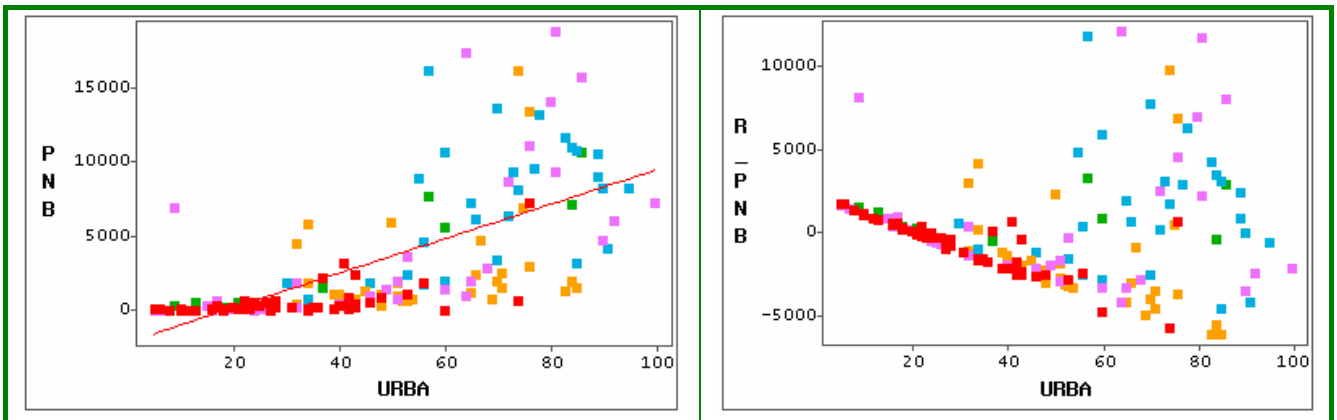


Figure 3.3 Effectifs par Continent

Sur la Figure 3.4 a, sont représentées la variable PNB, Produit national brut en fonction du taux d'urbanisation URBA, et la droite de régression, pour 173 pays.
Bien que les résultats statistiques soient significatifs (R^2=0.4358, F=122.06, *p value* <0.0001, T de Student = 11,05, *p value* <0.0001) cette analyse n'est pas satisfaisante.
La supposition de liaison linéaire n'est pas vérifiée, confirmé par le graphique des résidus (figure 3.3 b).

La variance des erreurs n'est pas constante, elle augmente en fonction de URBA. Il y a un effet d'entonnoir caractéristique de l'**hétéroscédasticité** des erreurs.



a) régression linéaire PNB en fonction de URBA

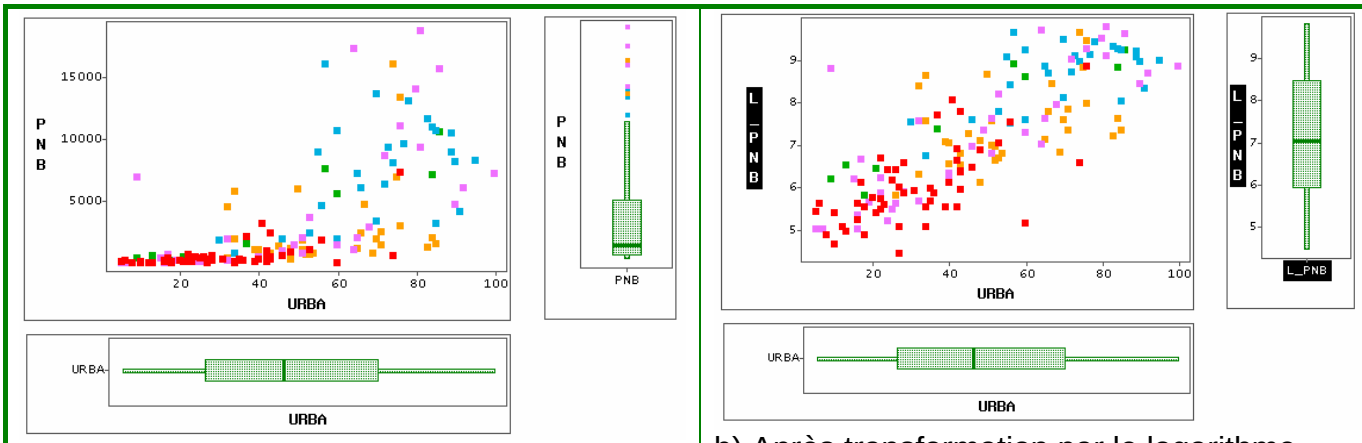
b) graphique des résidus R_PNB en fonction de URBA

Figure 3.4

Les couleurs des points observations correspondent aux couleurs des continents de la Figure 3.3.
On peut noter grâce aux couleurs que l'Afrique (point rouge) se différencie totalement de l'Europe (point bleu), on a à faire à des mélanges de population. Cependant pour notre démonstration sur les transformations on passera sous silence cette remarque.

Pour dilater les valeurs faibles et en même temps compresser les valeurs élevées du PNB on transforme la variable en son logarithme.

En SAS/INSIGHT il suffit de cliquer sur le nom de la variable PNB sur le graphique et de demander par le menu ➔ Edit # Variables # Log, la transformation de la variable PNB en son logarithme Log (L_PNB). Tous les affichages sont modifiés (figures 3.5 a et b).



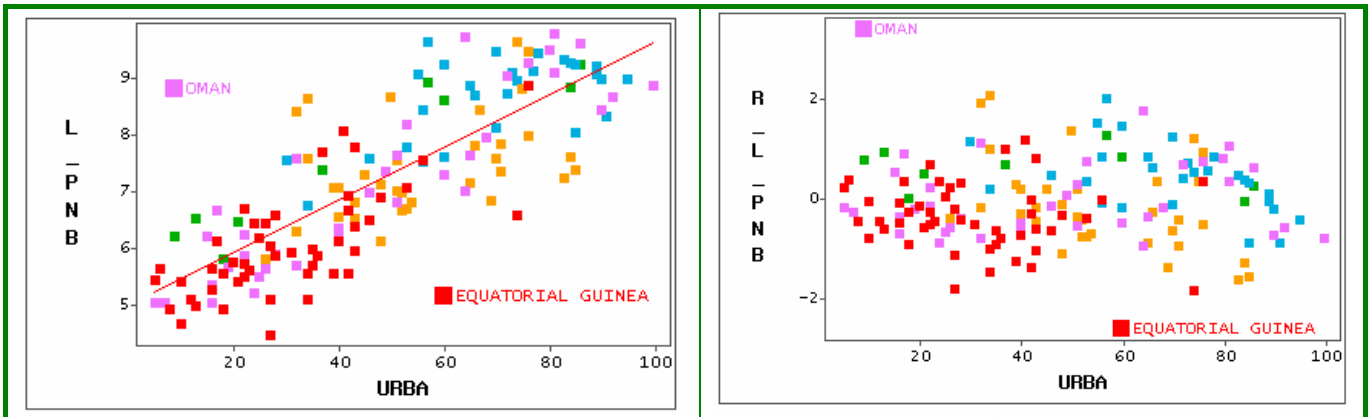
a) Avant transformation

b) Après transformation par le logarithme

Figure 3.5 Liaison entre le taux d'urbanisation et le Produit National Brut

La transformation a eu un double effet (figure 3.5). D'une part elle a symétrisé la distribution du PNB, visible sur les Box-plots, et d'autre part elle a linéarisé la liaison entre Log(PNB) et URBA

Les indicateurs statistiques de la régression sont nettement améliorés. (R^2=0.6468, F=289.39, *pvalue*<0.0001, T de Student = 34,05, *pvalue* <0.0001).



a) régression Log(PNB) fonction de URBA

b) Graphique des résidus fonction de URBA

Figure 3.6

Les résidus (figure 3.6 b) se trouvent maintenant bien repartis dans la bande (-2, +2), à l'exception de 2 pays atypiques, *Oman* et *Equatorial Guinea*, sur lesquels il faut s'interroger.

Echelle de Tukey – Ladder of Power

Les transformations de variables occupent une place importante dans la littérature. On a vu précedemment, que la transformation Log du PNB, permettait de linéariser la liaison entre Y et X. Pour choisir une transformation appropriée, J.W. TUKEY (1977) a proposé ce qu'on appelle maintenant l'échelle de transformation de TUKEY[15].

4	Y^4	Monter l'échelle
3	Y^3	
2	Y^2	
1	Y^1	brut
1/2	$\sqrt{Y}$	
0	$\mathrm{Log}(Y)$	Descendre l'échelle
-1/2	$-1/\sqrt{Y}$	
-1	$-1/Y$	
-2	$-1/Y^2$	

Echelle de Tukey

Selon la forme des courbes définies par les points (Xi,Yi) on pourra soit monter l'échelle, c'est-à-dire transformer X ou Y en ses puissances (Y^2, Y^3 etc.) soit descendre l'échelle, en prenant $\sqrt{Y}$, $\mathrm{Log}(Y)$, $-1/\sqrt{Y}$, $-1/Y$ etc.

Astuce de lecture proposée par J. Vanpoucke et E. Horber[16]

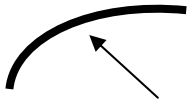
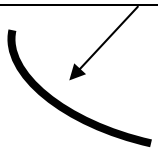
La forme des courbes du tableau 3.2 dessine un arc, et la flèche indique selon son orientation s'il faut monter ou descendre l'échelle.

Ainsi pour PNB en fonction de URBA, les données formaient une courbe d'allure n°3, il faut donc soit monter l'échelle en X, soit descendre l'échelle en Y. Comme la distribution de URBA est symétrique (voir figure 3.5 a), c'est plutôt sur PNB qu'il faut agir (distribution non symétrique, voir le Box plot de PNB).

Cette procédure raisonnée permet d'éviter d'agir complètement par « essais-erreurs », cependant pour trouver le bon choix on n'évite pas de faire quelques essais, grandement facilités par l'interactivité de SAS/INSIGHT.

[15] TUKEY utilise le mot « re-expression » et non le mot « transformation » d'une variable.
[16] JACQUES VANPOUCKE de l'Université Sabatier de Toulouse et EUGÈNE HORBER de l'Université de Genève, sont les fondateurs de l'Association MIRAGE, **M**ouvement **I**nternational pour le développement de la recherche en **A**nalyse **G**raphique et **E**xploratoire, http://www.unige.ch/ses/sococ/mirage/

Tableau 3.2 Transformations appropriées selon la forme des courbes

Forme de Courbes	Description de l'action	Transformation sur X	Transformation sur Y
	Descendre l'échelle en X ou Monter l'échelle en Y	$\sqrt{X}$, Log(X), $-1/\sqrt{X}$ etc.	Y^2, Y^3 etc.
	Descendre l'échelle en X ou Descendre l'échelle en Y	$\sqrt{X}$, Log(X), $-1/\sqrt{X}$ etc.	$\sqrt{Y}$, Log(Y), $-1/\sqrt{Y}$ etc.
	Monter l'échelle en X ou Descendre l'échelle en Y	X^2, X^3 etc.	$\sqrt{Y}$, Log(Y), $-1/\sqrt{Y}$ etc.
	Monter l'échelle en X ou Monter l'échelle en Y	X^2, X^3 etc.	Y^2, Y^3 etc.

Pour vous familiariser avec les transformations aller voir cet applet sur internet http://noppa5.pc.helsinki.fi/opetus/sd/sdt0.html. Développé par JUHA PURANEN de l'Université de Helsinski, cet applet montre l'effet de différentes transformations sur la liaison entre 2 variables, Y= Distance de freinage et X=Vitesse du véhicule, dont tout conducteur sait que la liaison n'est pas linéaire.
Il montre également la technique de lissage de courbes par LOWESS.

3.1.4. Méthode non paramètrique du LOWESS

La méthode non paramétrique du LOWESS - **Lo**cally **WE**ighted **S**moothing **S**catter-de Cleveland (1979, 1993, 1994) permet de lisser une courbe. Cette technique combine une technique de régression linéaire ou polynomiale locale avec la flexibilité de la régression non linéaire. Le lissage obtenu permet à l'œil de repérer des points atypiques, de voir d'éventuelles structures, de détecter des non linéarités etc.
Le principe repose sur des régressions locales définies sur des fenêtres glissantes. Chaque point observation est estimé, par une droite éventuellement un polynome, à partir des points de son voisinage, situés dans une fenêtre. Chaque point du voisinage est pondéré en fonction de sa distance au point estimé.
Les paramètres sur lesquels on peut agir sont la largeur de la fenêtre (Bandwith) et la fonction de pondération des points. Le principe est analogue à celui des moyennes mobiles, plus la largeur de la fenêtre est grande plus la série sera lissée. Pour la fonction de poids, on utilise généralement la fonction tri-cube.

$$W(x) = (1 - |x|^3)^3 \text{ pour } |x| < 1$$
$$W(x) = 0 \quad \text{pour } |x| >= 1$$

SAS/INSIGHT permet d'utiliser cette technique. L'exemple[17] porte sur les températures quotidiennes maximum de Melbourne de janvier 1981 à décembre 1990 (Source : Australian Bureau of Meteorology, 3650 observations).

Dans le menu Fit de la régression, on demande une régression simple de la variable *Température* en Y en fonction de la variable *Date* en X. Lorsque l'analyse de régression est affichée, on choisit le menu ➔ **Curves # Loess**

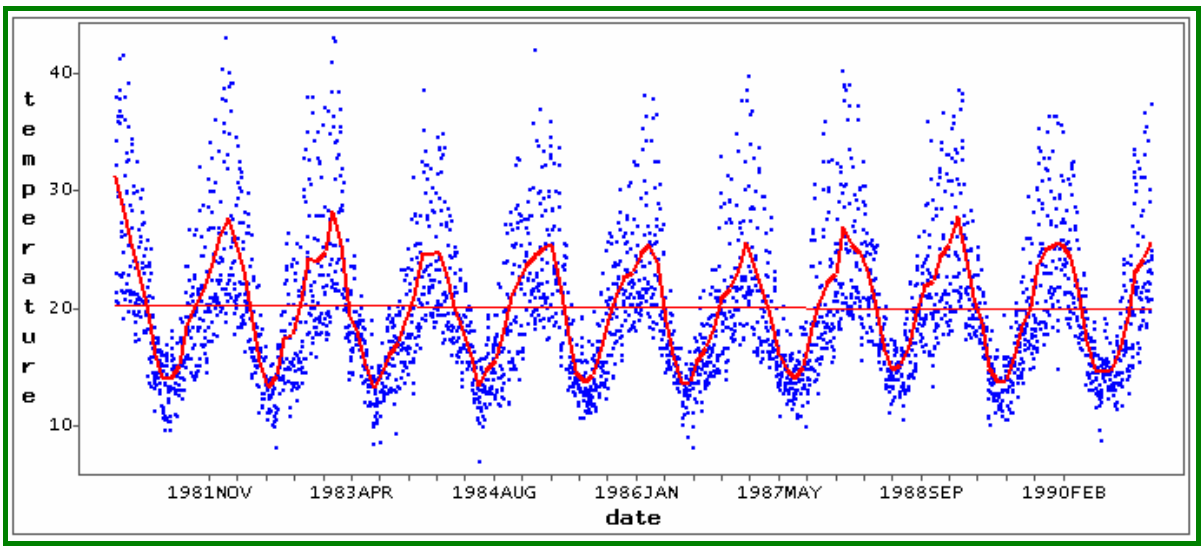


Figure 3.6 Technique du LOWESS appliquée à des données de températures

La droite de régression est strictement horizontale avec une température moyenne de 21°04 sur les 10 années (figure 3.6).

Model Equation					
temperature	=	21.0418	-	0.0001	date

La technique du LOWESS (parfois dénommée LOESS) permet le lissage de la série. Elle nécessite beaucoup de calculs, ce qui ne pose plus de problèmes avec les ordinateurs actuels si la technique est bien programmée. C'est une des techniques modernes les plus attractives puisqu'elle ne nécessite pas de préciser la forme d'un modèle, elle laisse « **parler les données** » Cette technique est très utile dans la phase exploratoire des données mais elle a son revers, elle ne fournit pas de fonction analytique comme la régression linéaire.

[17] Les données proviennent de la banque de données de Rob J . Hyndman sur des séries temporelles http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/

Tableau 3.3

Loess Fit										
Curve	Type	Weight	N_Interva	Metho	Alpha	K	DF	R-Square	MSE	MSE(GCV)
	Line	Tri-Cu	128	GCV	0.0194	70	123.56	0.5419	17.6874	18.3071

En cliquant sur le curseur Alpha (paramètre du LOWESS lié à la largeur de bande - bandwith) on peut agir sur la largeur de la fenêtre et voir l'effet sur le filtrage. Avec un coefficient Alpha de 0.005, on fait apparaître un lissage moins régulier.

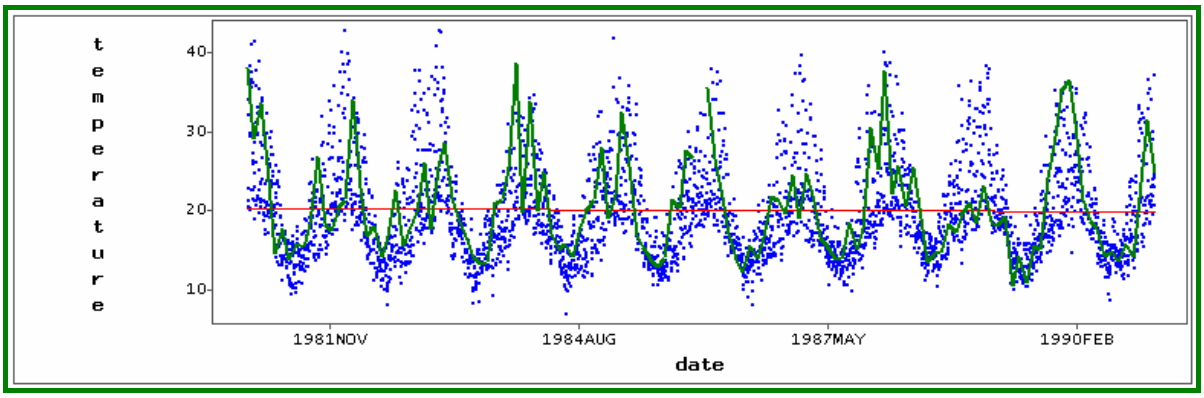


Figure 3.7 après modification du paramètre de filtrage Alpha

Pour connaître la technique du LOWESS programmée dans SAS, il suffit de cliquer sur un mot clé (par exemple Alpha –voir Tableau 3.3) et de demander l'aide en ligne par le menu
➔ **Help # Help on Selection**.

Il existe aussi dans SAS une procédure LOESS dans le module SAS/ETS.

3.2. Exemples en régression multiple

3.2.1. Y « expliquée » par la corrélation entre deux régresseurs

Cet exemple, montre l'influence sur le coefficient de détermination lorsque les régresseurs sont corrélés. Les données sont issues du cours de Georges Monette de York University.

```
Data HWH;
input Weight Height Health;
cards;
68 94 120
137 114 60
94 104 123
121 107 94
100 118 104
93 91 117
76 123 139
102 73 100
122 112 91
89 78 91
```

```
69 61 103
123 150 131
33 60 128
207 193 107
135 153 141
;
proc reg;
 ModHW: model health=weight;
 ModHH: model health=Height;
 ModHWH: model health=weight Height;
run ;
```

- Pour le premier modèle, régression simple de HEALTH en fonction de WEIGHT, le coefficient de détermination vaut **0.0738. Aucune liaison entre santé et poids.**
- Pour le deuxième modèle, régression simple de HEALTH en fonction de HEIGHT, le coefficient de détermination vaut **0.0367. Aucune liaison entre santé et taille.**
- Pour le troisième modèle du tableau 3.4, donnant les résultats de la régression multiple de HEALTH en fonction de WEIGHT et HEIGHT, le coefficient de détermination vaut **0.6551.** La statistique F et les T de student sont tous significatifs. Si on s'arrête à ces seules indications le modèle explique 65 % de la variation de Health, voir le tableau 3.4.

Tableau 3.4 Régression de Health en fonction de WEIGHT et HEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4295.20012	2147.60006	11.39	0.0017
Error	12	2261.73322	188.47777		
Corrected Total	14	6556.93333			

Root MSE	13.72872	R-Square	0.6551
Dependent Mean	109.93333	Adj R-Sq	0.5976
Coeff Var	12.48822		

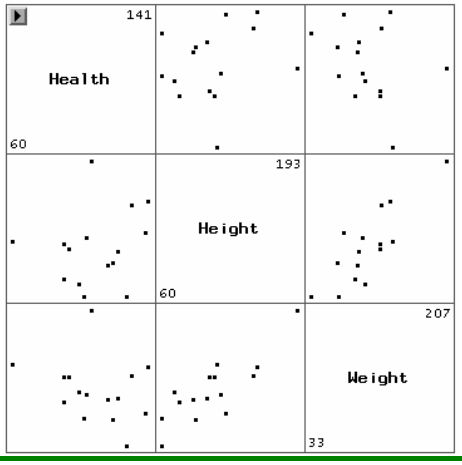
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	101.08315	11.55217	8.75	<.0001
Weight	1	-0.77406	0.16689	-4.64	0.0006
Height	1	0.82603	0.18369	4.50	0.0007

Cette fausse explication de la santé par le Poids et le Taille est due à la corrélation entre les régresseurs (0.8357). Ce qui est visible sur la matrice de corrélation et sur la matrice de scatter plot.

Matrice de corrélation et Scatter Plot

Correlation Matrix	Health	Height	Weight
Health	1.0000	0.1916	-0.2717
Height	0.1916	1.0000	0.8357
Weight	-0.2717	0.8357	1.0000



3.2.2. Instabilité des coefficients de la régression, en cas de multicolinéarité

Exemple sur données réelles

Cet exemple est de D. LADIRAY [18]. On dispose de 44 observations et on cherche à expliquer le taux d'urbanisation, variable URBA, en fonction de 10 variables régresseurs, POP87 à ESPER.

Tableau 3.5 Tableau des données

OBS	URBA	POP87	NAT	MORT	ACCR	DOUB	FERTI	MORTI	AGE15	AGE65	ESPER	CCR_CAL
1	81	0.4	32	5	2.8	25	4.6	32.0	41	2	67	2.7
2	53	0.7	20	9	1.1	63	2.5	12.0	25	11	74	1.1
3	79	0.6	47	8	4.0	18	7.4	59.0	50	3	64	3.9
4	68	17.0	46	13	3.3	21	7.2	80.0	49	4	62	3.3
5	90	4.4	23	7	1.7	41	3.1	12.3	33	9	75	1.6
6	60	3.7	45	8	3.7	19	7.4	54.0	51	3	67	3.7
7	80	1.9	34	3	3.2	22	4.4	19.0	40	1	72	3.1
8	80	3.3	30	8	2.2	32	3.8	52.0	38	5	65	2.2
9	9	1.3	47	14	3.3	21	7.1	117.0	44	3	52	3.3
10	86	0.3	34	4	3.0	23	5.6	42.0	34	2	69	3.0
11	72	14.8	**39**	7	3.1	22	6.9	79.0	37	2	63	3.2
12	49	11.3	47	9	3.8	18	7.2	59.0	49	4	63	3.8
13	46	51.4	30	9	2.1	33	4.0	92.0	36	4	62	2.1
14	81	1.4	30	4	2.6	27	5.9	38.0	30	1	68	2.6
15	15	6.5	53	19	3.4	20	7.8	137.0	49	3	47	3.4
16	40	2.4	47	17	3.0	23	7.3	135.0	48	3	48	3.0
17	16	14.2	48	22	2.6	27	7.6	182.0	46	4	39	2.6
18	13	107.1	44	17	2.7	26	6.2	140.0	44	4	50	2.7
19	5	1.5	38	18	2.0	34	5.5	142.0	40	3	46	2.0
20	25	800.3	33	12	2.1	33	4.3	101.0	38	4	55	2.1
21	51	50.4	45	13	3.2	21	6.3	113.0	44	3	57	3.2
22	26	0.2	48	10	3.8	18	7.1	68.0	45	2	51	3.8
23	7	17.8	42	17	2.5	28	6.1	112.0	41	3	52	2.5
24	28	104.6	44	15	2.9	24	6.6	125.0	45	4	50	2.9
25	22	16.3	25	7	1.8	38	3.7	29.8	35	4	70	1.8
26	64	0.2	30	4	2.6	26	3.6	12.0	38	3	62	2.6
OBS	URBA	POP87	NAT	MORT	ACCR	DOUB	FERTI	MORTI	AGE15	AGE65	ESPER	CCR_CAL

[18] Ladiray D.(1990) *Autopsie d'un résultat: L'exemple des procédures Forecast, X11, Cluster.* Club SAS 1990

27	24	38.8	34	13	2.1	33	4.4	103.0	39	4	53	2.1
28	12	0.7	48	23	2.5	28	5.8	183.0	35	3	40	2.5
29	22	174.9	31	10	2.1	33	4.2	88.0	40	3	58	2.1
30	11	6.5	**39**	18	2.1	33	4.7	160.0	35	3	43	2.1
31	16	3.8	41	16	2.5	28	5.8	122.0	43	3	50	2.5
32	32	16.1	31	7	2.4	28	3.9	30.0	39	4	67	2.4
33	40	61.5	35	7	2.8	25	4.7	50.0	41	3	65	2.8
34	100	2.6	17	5	1.1	61	1.6	9.3	24	5	71	1.2
35	17	53.6	29	8	2.1	33	3.5	57.0	36	3	63	2.1
36	19	62.2	34	8	2.6	27	4.5	55.0	40	4	63	2.6
37	32	1062.0	21	8	1.3	53	2.4	61.0	28	5	66	1.3
38	92	5.6	14	5	0.9	77	1.6	7.5	24	7	75	0.9
39	76	122.2	12	6	0.6	124	1.8	5.5	22	10	77	0.6
40	64	21.4	30	5	2.5	28	4.0	33.0	39	4	65	2.5
41	65	42.1	20	6	1.4	51	2.1	30.0	31	4	67	1.4
42	97	0.4	23	6	1.7	41	3.7	12.0	34	8	68	1.7
43	51	2.0	37	11	2.6	26	5.1	53.0	42	3	62	2.6
44	67	19.6	17	5	1.2	59	1.8	8.9	30	5	73	1.2

Dans le tableau 3.5, deux valeurs de la variable NAT (taux de natalité) pour OBS=11 et OBS=30 sont légèrement modifiées (39 est remplacé par 40)
Les régressions effectuées avant et après modifications donnent les résultats suivants pour les coefficients de régression :

Tableau 3.6

Résultat 1 (valeur 39) **Avant**		**Résultat 2** (valeur 40) **Après**	
URBA =		**URBA=**	
25.541		20.689	
-0.026	POP87	-0.026	POP87
-6.661	**NAT**	**-4.047**	**NAT**
+2.681	**MORT**	**-0.005**	**MORT**
+64.506	**ACCR**	**+39.832**	**ACCR**
+0.019	DOUB	+0.015	DOUB
+7.834	FERTI	+7.307	FERTI
+0.101	MORTI	+0.128	MORTI
-1.132	AGE15	-1.157	AGE15
+2.709	AGE65	+2.848	AGE65
+0.910	ESPER	+0.969	ESPER

Les résultats "Avant" et "Après" (tableau 3.6) sont particulièrement instables pour les estimations des coefficients des 3 variables, NAT (taux de natalité), MORT (taux de mortalité) et ACCR (taux d'accroissement de la population).

Explication
Ces 3 variables **ne sont pas indépendantes**, elles sont liées entre elles par une relation quasi-linéaire ACCR= (NAT-MORT)/10. Dans le tableau 3.5, on peut comparer ACCR avec la variable *ACCR_CAL*, en dernière colonne, calculée avec la formule exacte.
Lors de l'inversion de la matrice X'X, il y a une valeur propre qui est presque nulle. Conséquence une légère perturbation des données entraîne de grands changements dans les estimations des paramètres.

***Exemple sur données avec modèle théorique connu* et régresseurs corrélés**

Cet exemple est de T. Foucart (2007). Pour étudier l'effet des corrélations entre les régresseurs sur les estimations des paramètres, T. Foucart a généré 100 observations d'un **vrai modèle théorique** :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Les vraies valeurs des paramètres du modèle théorique sont :

$$\beta_1 = 0 \qquad \beta_1 = \beta_2 = 0.5 \qquad \beta_3 = \beta_4 = -0.5$$

Les 4 régresseurs X_1 à X_4 suivent des lois normales centrées et réduites.
L'erreur ε suit une loi normale $N(0, \sigma^2)$,
On impose de plus des contraintes sur la matrice des corrélations entre les régresseurs.

Tableau 3.7 Corrélations imposées

	X_1	X_2	X_3	X_4
X_1	1			
X_2	0.5	1		
X_3	0.5	0.5	1	
X_4	-0.5	0.4	0.3	1

Et on impose la valeur du coefficient de détermination $R^2 = 0.5$.

Les données générées RIDGE1 sont disponibles sur le site de T. Foucart[19].

Résultats avec SAS/INSIGHT

- la matrice de corrélation
-

Menu ➔ **Analyze # Mutivariate** avec les 5 variables X1, X2, X3, X4, Y dans le rôle Y

On la trouve dans le (tableau 3.8) ci-dessous.

[19] http://foucart.thierry.free.fr/StatPC/.

Tableau 3.8 Statistiques univariées et matrice de corrélation

X1 X2 X3 X4 Y

Univariate Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
X1	100	-2.22E-16	1.0050	-2.9987	2.8081
X2	100	0	1.0050	-2.9509	2.7650
X3	100	4.1633E-17	1.0050	-2.6484	2.6579
X4	100	0	1.0050	-2.9459	2.0666
Y	100	-0.1650	1.5233	-4.1021	3.4373

Correlation Matrix

	X1	X2	X3	X4	Y
X1	1.0000	0.5000	0.5000	-0.5000	0.5404
X2	0.5000	1.0000	0.5000	0.4000	0.2161
X3	0.5000	0.5000	1.0000	0.3000	-0.1065
X4	-0.5000	0.4000	0.3000	1.0000	-0.4906
Y	0.5404	0.2161	-0.1065	-0.4906	1.0000

On vérifie sur le tableau 3.8 que la matrice de corrélation a bien les valeurs imposées du Tableau 3.7.

- <u>la matrice de scatter plot</u>

Menu ➔ **Analyze# Scatter Plot** avec les 5 variables (X1, X2, X3, X4, Y) dans le rôle X et les mêmes dans le rôle Y

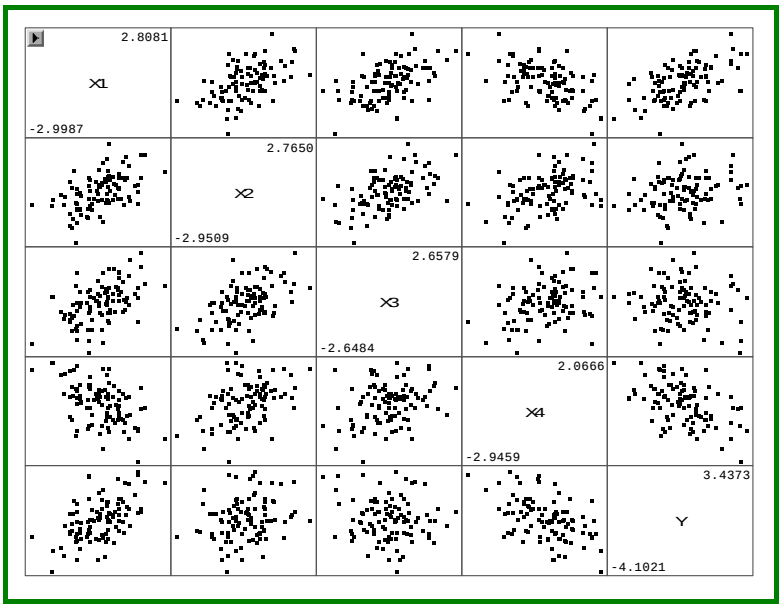


Figure 3.6 Matrice de diagrammes de dispersion

La matrice de diagrammes de dispersion permet de repérer les liaisons entre les régresseurs et la variable Y. Y est lié à X1, X2, avec un coefficient positif, et Y est lié à X3, X4 avec un coeffcient négatif.

- <u>La régression linéaire</u>

Menu ➔ **Analyze#Fit** avec les 4 variables X1, X2, X3, X4, dans le rôle de X et Y dans le rôle de Y

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	4	112.7185	28.1796	22.88	<.0001
Error	95	117.0120	1.2317		
C Total	99	229.7305			

Summary of Fit			
Mean of Response	-0.1650	R-Square	0.4907
Root MSE	1.1098	Adj R-Sq	0.4692

Parameter Estimates					
Variable	DF	Estimate	Std Error	t Stat	Pr >|t|
Intercept	1	-0.1650	0.1110	-1.49	0.1405
X1	1	1.6339	0.8739	1.87	0.0646
X2	1	-0.1482	0.5659	-0.26	0.7940
X3	1	-1.0375	0.4153	-2.50	0.0142
X4	1	0.4439	0.7848	0.57	0.5730

Avec la méthode des moindres carrés, le coefficient de détermination R^2=0.4907 est bien calculé (le théorique vaut 0.50), par contre les coefficients sont très différents de ceux du modèle théorique.

Les estimations MCO sont respectivement : 1.6339, -0.1482, -1.0375, 0.4439 au lieu des vraies valeurs : 0.5, 0.5, -0.5, -0.5. Mêmes les signes ne sont pas respectés.

Conclusion

Les conséquences des colinéarités entre les variables régresseurs sont les suivantes :

- Les coefficients de régression sont instables ;
- Leur signe peuvent changer (positif ←→ négatif) rendant les interprétations fausses, ce qui a de graves conséquences lors de la recherche des effets d'une variable régresseur ;
- Les variances des estimateurs sont élevées.

La technique de la régression bornée (Ridge Regression) a été proposée dans les années 1970, pour pallier ces inconvénients. On en trouvera un exemple au chapitre 4 (paragraphe 4.4.5). L'article de T. FOUCART (2007) « Evaluation de la régression bornée »[20], montre que là encore on ne peut systématiquement y recourir. Il faut bien connaître les données et le domaine pour en faire bon usage.

3.3. Conditions d'utilisation de la régression, les diagnostics

Les différents exemples présentés dans ce chapitre montrent l'importance des analyses et diagnostics effectués avant et après les premiers traitements.

Pour réaliser les tests d'hypothèses de la régression on a supposé, en pure théorie, que le modèle linéaire postulé est correct, que les suppositions, *a priori* sur les variables et sur les erreurs, sont vraies. Ces suppositions sont nécessaires pour

[20] Site http://foucart.thierry.free.fr/colreglin/Regression_bornee.pdf,

définir les tests comme on l'a vu au Chapitre 1, car on ne peut calculer les variances que dans des cas gaussiens et sous certaines conditions.

Suppositions sur les variables

- Les observations Y_i sont supposées indépendantes
- Les variables X_j sont non aléatoires.

Suppositions sur les erreurs

- les erreurs sont d'espérance nulle ce qui est vérifié par construction si la constante β_0 existe dans le modèle.
- les erreurs sont de variance constante
- les erreurs suivent une distribution normale
- les erreurs sont indépendantes

Les erreurs sont inconnues, elles sont approchées par les résidus, **si le modèle est correct.**

Après examen des résidus, on peut conclure: les suppositions semblent ou ne semblent pas être violées. Ce qui ne signifie pas que les suppositions soient correctes. Cela veut dire que sur la base des données que l'on a étudiée, on n'a aucune raison de dire que les suppositions sont fausses.

3.3.1. Modèle Inadapté

C'est par l'examen des résidus que l'on peut voir si le modèle postulé est vraisemblablement correct, ou s'il est **inadapté**.
Les résidus contiennent à la fois des erreurs de mesure et des erreurs de spécification du modèle, comme des variables omises, ou des liaisons non linéaires.
On peut avoir des tests satisfaisants, de bonnes précisions sur les estimateurs des paramètres, alors que le modèle est inadapté à l'étude.

En dehors de certaines visualisations il n'y a que le bon sens et la **pré-connaissance du problème** qui permettent de repérer l'inadéquation du modèle aux données.

3.3.2. L'influence de certaines données, les données atypiques -Outliers-

Certaines données atypiques peuvent fausser les résultats. Les visualisations graphiques permettent parfois de les identifier.

Le livre de BELSLEY, KUH et WELSH (BKW) a popularisé une méthode rigoureuse de recherche des observations influentes. L'option INFLUENCE de Proc REG permet cette analyse.
On peut alors être amené à retirer ces points atypiques des analyses, ou à procéder à des techniques « **robustes** » (voir ROUSSEEUW et al. (2003)). La procédure ROBUSTREG de SAS disponible en V9 reprend ces techniques.

3.3.3. Corrélation et colinéarité entre les régresseurs

La colinéarité est un **gros problème,** lot quotidien du statisticien praticien lorsqu'il analyse des données réelles, principalement en sciences économiques et sociales.
C'est également le trio "BKW" qui a proposé des indicateurs de détection de colinéarités.
Les options TOL, VIF et COLLIN, COLLINOINT de Proc REG sont des aides aux diagnostics de colinéarité.

Tous ces compléments à la régression, représentations graphiques, indicateurs techniques de BKW, etc., nécessitent de faire appel aux nombreuses options de Proc REG, qui seront présentées au chapitre 4 « Validation d'une régression ».

4. Validation d'une régression

Dans ce chapitre, on présente les différents éléments nécessaires à la validation d'une régression, c'est-à-dire la vérification des suppositions de base du modèle, l'étude de la robustesse au niveau des observations (détection des observations influentes et atypiques) et au niveau des variables (colinéarités, choix d'un sous-ensemble de régresseurs).

Les sorties de SAS-version 9 illustrant ce chapitre sont réalisés avec le même exemple que pour le chapitre 2 (§2.3.1), issu du livre de Tomassone et *al*. (1992).

4.1. Introduction

Un principe de base doit être appliqué : explorer les données par des graphiques et/ou des calculs numériques.

Les calculs préalables des caractéristiques des variables Y et des Xj, ainsi que des histogrammes et des tracés Box-Plot de ces différentes variables permettent en effet de mettre en évidence des problèmes. Bien sûr le calcul des corrélations entre Y et les Xj est nécessaire.

Il faut étudier les variables, et donc en particulier faire des graphiques, par exemple de Y contre les régresseurs Xj pour contrôler la linéarité des liaisons. Pour cela, on utilisera la procédure GPLOT, ou le menu « Scatter-Plot » de SAS/INSIGHT.
Dans PROC REG, l'instruction PLOT permet de faire des graphiques des variables entrant dans la régression, mais aussi des variables créées comme les valeurs résiduelles et ajustées.

De plus, une Analyse en Composantes Principales des régresseurs, avec Y en variable supplémentaire, peut aussi être utile pour visualiser comment Y se reconstruit à partir de l'ensemble des X.

4.1.1. Modèle et notations

Si on postule un modèle avec n observations (i) et p variables régresseurs (j) et une constante, on note :

- **Y** réponse ou variable dépendante
- **X** matrice des variables régresseurs
- β coefficients de régression
- ε erreurs.

Le modèle linéaire s'écrit : $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_p X_{ip} + \varepsilon_i$.

D'où l'ajustement:

$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + ... + b_p X_{ip}$ et les résidus $e_i = Y_i - \hat{Y}_i$.

Suppositions sur les erreurs : ε_i sont des aléas, indépendants, d'espérance nulle, de variance constante, et de même loi (cf. chapitre 1).
→on dit « IID avec loi normale $N(0,\sigma^2)$».

4.1.2. Problèmes à étudier

Vérification des suppositions sur les erreurs ;

Robustesse de la régression : détection des observations influentes, et de la colinéarité des régresseurs ;

Choix d'un sous-ensemble de régresseurs.

4.2. Vérification des suppositions de base sur les erreurs

Les suppositions sur les erreurs (inconnues) doivent être vérifiées à partir de leurs observations (les résidus).

4.2.1. Espérance nulle

Il faut vérifier que les résidus sont de moyenne nulle. Or les résidus construits par les moindres carrés sont centrés par construction, si la constante est dans le modèle (ce que l'on suppose ici).

4.2.2. Indépendance

Il faudrait vérifier que le vecteur des résidus forme un échantillon tiré de n variables aléatoires indépendantes

On obtient $\hat{Y}$ par projection du vecteur Y sur le sous-espace engendré par les régresseurs : il en résulte que les n composants du vecteur e des résidus, $e_i = Y_i - \hat{Y}_i$, sont reliés par des relations ainsi que le montre la figure 4.1 :

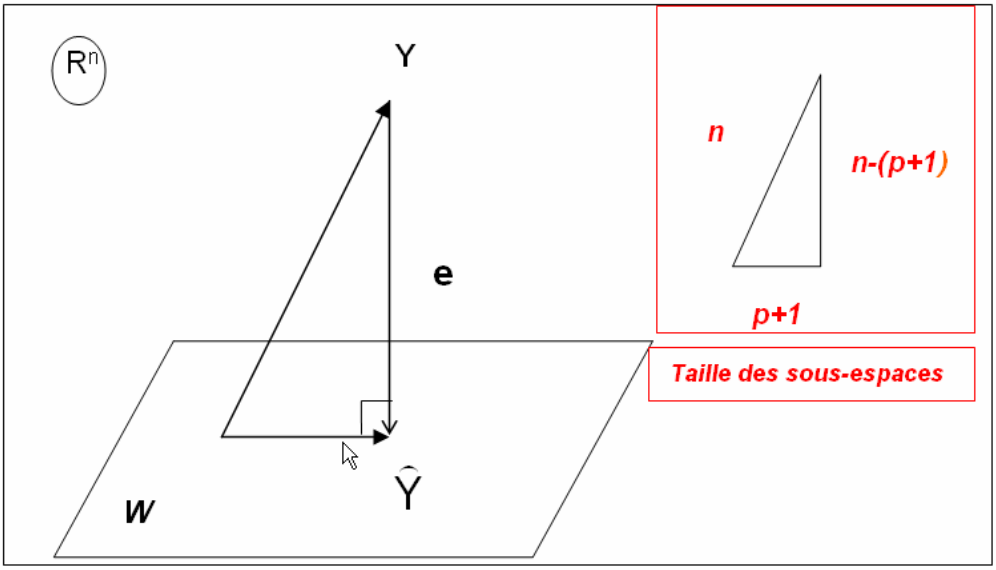


Figure 4.1 : Projection de Y dans l'espace des régresseurs

Dans la représentation géométrique dans l'espace R^n, le vecteur des n résidus est situé dans le sous-espace orthogonal à celui des régresseurs ; celui-ci étant de dimension (p+1), le vecteur des résidus est alors situé dans un espace de dimension (n-(p+1)).
L'indépendance n'a donc de sens que si n est grand par rapport à p.

Remarque :
De façon générale, pour tester l'indépendance, on met en œuvre des tests non paramétriques (qui ne sont proposés dans SAS) basés sur des séquences : séquence des signes des différences successives (e_{i+1} - e_i), ou séquence des signes des différences à la médiane (e_i - Mediane).

Cas particulier où les observations sont apparentées (cas des chroniques) :

Alors le test de Durbin-Watson permet de vérifier si le résidu en i est non-corrélé au résidu en (i+1) : on parle d'auto-corrélation d'ordre 1. Il est obtenu par l'option DW de l'instruction MODEL de Proc REG.

On calcule ainsi le coefficient de Durbin-Watson à partir des résidus $e_i = Y_i - \hat{Y}_i$,

$$DW = \frac{\sum_i (e_{i+1} - e_i)^2}{\sum_i e_i^2}$$

En notant $\rho = \left(\frac{\sum e_{i+1}.e_i}{\sum e_i^2} \right)$, si les résidus forment un *processus autorégressif d'ordre 1,*
c'est-à-dire suivent le modèle $e_{i+1} = \rho.e_i + \eta_i$, alors DW vaut à peu près $2(1-\rho)$,

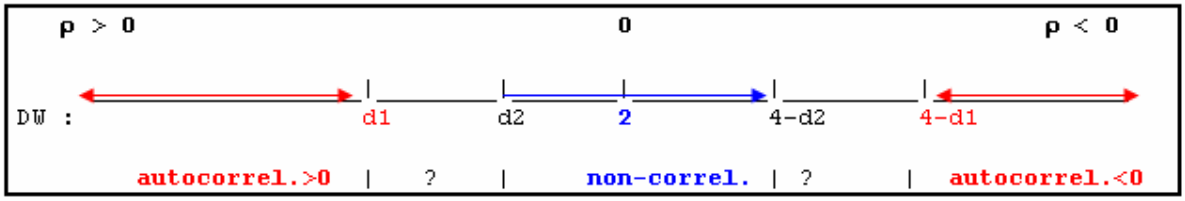
où $\boxed{DW \cong 2.\left(1-\frac{\sum e_{i+1}.e_i}{\sum e_i^2}\right)}$.

Liens entre les valeurs ρ et DW:

Si $0 < \rho < 1 \Rightarrow$ DW compris entre 0 et 2
Si $0 > \rho > -1 \Rightarrow$ DW compris entre 2 et 4

S'il n'y a pas d'auto-corrélation d'ordre 1 $\Leftrightarrow \rho$ proche de 0, donc DW proche de 2.

Il existe des tables dites de Durbin-Watson permettant de tester l'absence d'auto-corrélation d'ordre 1 en fonction du niveau de confiance α, et de n (nombre d'observations) et p (nombre de variables). On y lit deux valeurs d1 et d2:



Les causes de l'auto-corrélation sont une mauvaise spécification du modèle, ou l'absence d'une variable importante. *Les remèdes* sont soit de travailler sur les différences premières en Y c'est-à-dire (Y_i - Y_{i-1}), soit d'appliquer la méthode de Cochran-Orcutt (voir les livres spécialisés en Econométrie). PROC AUTOREG, du module SAS/ETS, réalise des régressions où le problème de l'auto-corrélation des résidus est résolu.

Remarque : Beaucoup de logiciels donnent systématiquement cette statistique DW, mais interpréter le test de Durbin-Watson sur les résidus n'a aucun sens si les données ne sont pas apparentées.

4.2.3. Egalité des variances (homoscédasticité)

Les graphiques des résidus contre les différents régresseurs Xj, permettent de visualiser si les résidus sont répartis dans une bande de valeurs horizontale autour de 0, c'est à dire s'il y a homoscédasticité. Sinon on peut alors détecter quelle est la variable responsable de l'hétéroscédasticité.

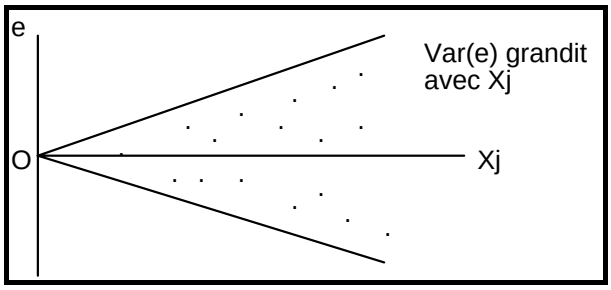


Figure 4.2 : Graphique « typique » des résidus contre Xj révélant une hétéroscédasticité

Ces graphiques peuvent être tracés par la procédure GPLOT si on a stocké les résidus dans une table, mais peuvent aussi être faits à l'intérieur de la procédure REG, à l'aide de l'instruction PLOT, avec «R. » comme nom de variable en ordonnée car le résidu est stocké en interne dans la variable ayant le nom R.
Dans SAS/INSIGHT, un graphique des résidus (dénommé R_Y) contre X est tracé si le modèle est une régression à un régresseur ; sinon c'est le tracé de R_Y contre l'estimation $\hat{Y}$ (dénommée P_Y) qui est tracé par défaut. Ces deux variables sont automatiquement créées dans la table SAS active. Il est possible de réaliser les autres graphiques à l'aide de la variable R_Y.

L'instruction MODEL de Proc REG possède une option SPEC pour tester s'il y a un problème d'hétéroscédasticité : l'hypothèse nulle « homoscédasticité » est testée à l'aide d'une statistique suivant une loi du Chi2 (cf. White H., (1980)). Le test est global, et donc en cas de rejet de H0, on ne sait pas quelle est la variable responsable de l'hétéroscédasticité.

D'autres tests, comme ceux de Goldfeldt et Quandt, ou de Breush et Pagan (voir les publications spécialisées en Econométrie, par exemple Green (2005)), permettent de mettre en évidence l'hétéroscédasticité dûe aux différentes variables. Mais ils sont assez lourds à mettre en œuvre, et ne sont pas faits par des procédures SAS.

Une méthode plus simple peut permettre d'avoir une idée préalable sur l'existence d'un problème. Celle-ci s'apparente au test de Chow(1960) pour une série chronologique, où on teste l'égalité des variances des résidus de 2 sous périodes de la chronique.
Ici, on trie les résidus selon les valeurs croissantes de la variable Xj suspecte (par PROC SORT), ensuite on partage le vecteur des résidus triés en 2 paquets (premiers, derniers) dont on calcule les variances. Puis on teste l'hypothèse nulle d'égalité de ces 2 variances c'est à dire la possibilité d'homoscédasticité, à l'aide de la procédure TTEST de SAS.

Quelques remèdes en cas d'hétéroscédasticité :

- transformer Y ou Xj par une fonction racine carrée, ou Log, ou carré, etc. pour « aplatir » les variances : l'échelle de Tukey donnée au chapitre 3 (§3.1.3), peut aider au choix de la transformation ;
- mettre en œuvre une régression pondérée avec l'instruction WEIGHT, en prenant comme poids $\frac{1}{\sqrt{f(X_j)}}$ si la variance σ^2 est une fonction connue f de X_j ;
- mettre en œuvre les moindres carrés généralisés, ce qui peut se faire par PROC GLM .

4.2.4. Normalité des erreurs

Supposition : Les ε_i sont indépendant, et suivent une loi $N(0,\sigma^2)$.

Cette supposition de normalité est nécessaire pour effectuer les tests sur les coefficients et les tests sur les sommes de carrés à l'aide des statistiques de Student ou de Fisher vues aux chapitres 1 et 2.

Comme tout test, le test d'adéquation à une loi nécessite l'indépendance. Or les résidus sont liés. Donc réaliser un test de normalité à une loi $N(0, \sigma^2)$ sur ces résidus n'a pas de sens.

Dans SAS, un tracé « QQ-Plot »[21] pour les résidus permet de vérifier graphiquement l'adéquation à la loi normale $(0,s^2)$ ou s^2 est estimé par MSE (Mean Square Error du modèle). Le QQ-Plot est obtenu dans la procédure REG avec l'instruction PLOT : on demande le tracé de R (variable interne des résidus) contre NQQ (variable interne contenant les quantiles de la loi normale) : voir l'exemple ci-dessous (§4.2.5). Dans SAS/INSIGHT, une fois que l'on a exécuté le modèle, on peut ajouter aux sorties standards un graphique QQ-Plot appelé « Residual Normal QQ » dans le menu Graphs (penser à cocher « Reference lines »dans le menu contextuel du graphique pour tracer la droite).

4.2.5. Exemple

On utilise les données « Processionnaire du pin » issu du livre de Tomassone et a*l.*.(1983), déjà traitées au chapitre 2. On se limite dans ce paragraphe au modèle Y = log = f(X1 X2 X4 X5), dont on verra que c'est un « bon » modèle.
On trouvera au §2.3.2 les caractéristiques des variables, la matrice de corrélation et les graphiques de dispersion des variables.

Modèle

```
proc reg data=libreg.chenilles; title 'Modèle Y = X1 X2 X4 X5
';
LOG : model LOG=X1 X2 X4 X5 ;
run ;
quit;
```

[21] On trouvera en annexe 5 le principe de construction des QQ-Plot pour l'adéquation à certaines lois.

Modèle Y = X1 X2 X4 X5

The REG Procedure
Model: LOG
Dependent Variable: log

Number of Observations Read	33
Number of Observations Used	33

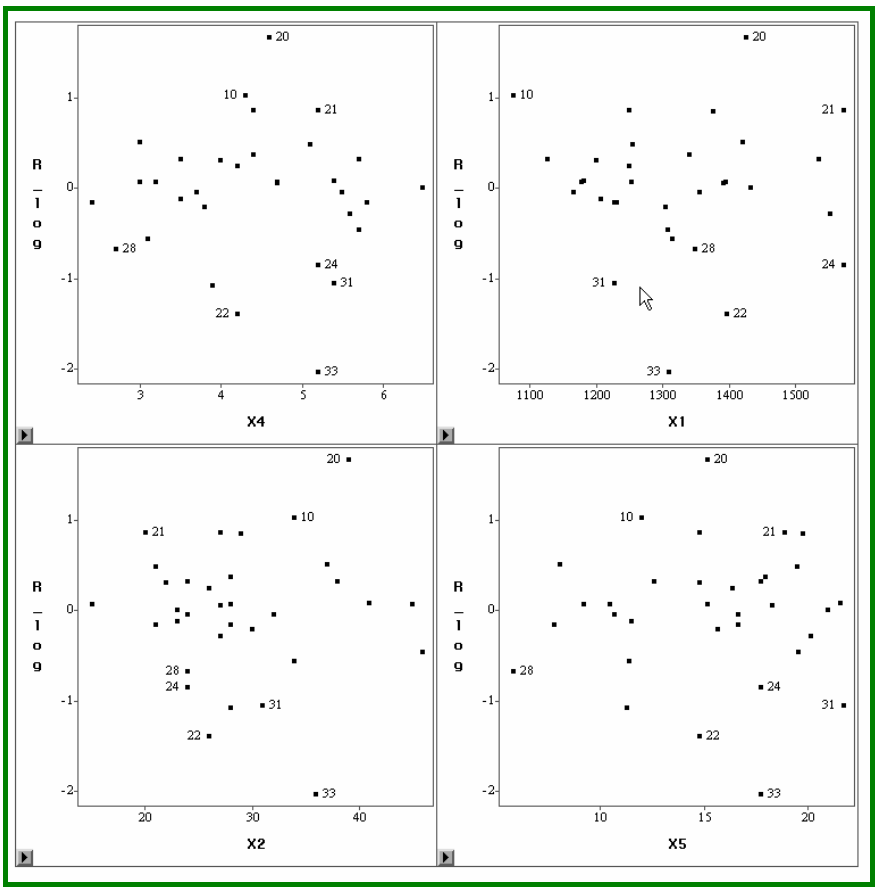
Analyse de variance					
Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	4	32.09265	8.02316	12.83	<.0001
Error	28	17.50338	0.62512		
Corrected Total	32	49.59603			

Root MSE	0.79065	R-Square	0.6471
Dependent Mean	-0.81328	Adj R-Sq	0.5967
Coeff Var	-97.21683		

Résultats estimés des paramètres					
Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	7.73214	1.48858	5.19	<.0001
X1	1	-0.00392	0.00115	-3.42	0.0019
X2	1	-0.05734	0.01939	-2.96	0.0062
X4	1	-1.35614	0.31983	-4.24	0.0002
X5	1	0.28306	0.07626	3.71	0.0009

Dessin des résidus contre les 4 régresseurs (avec SAS/INSIGHT)

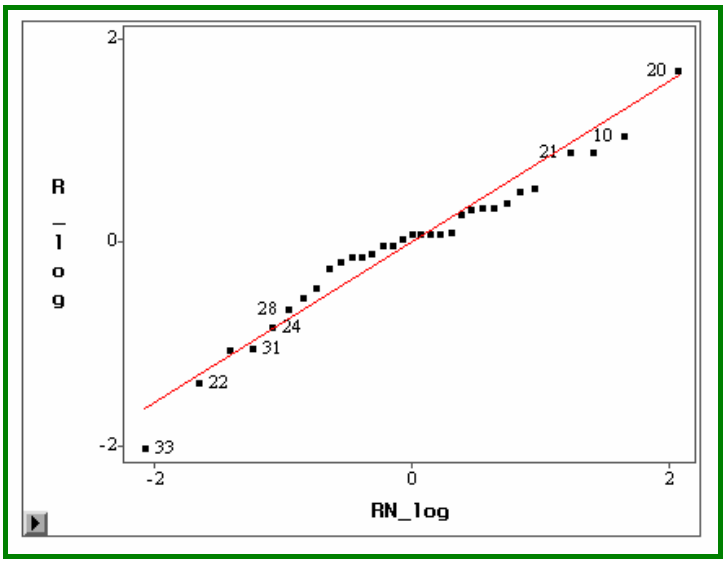
On y affiche le numéro de certaines observations, qui ont des résidus un peu grands, ou qui seront détectés plus loin comme « atypiques » (§4.3.11)



Remarque : instruction de tracé des résidus dans PROC REG

```
/* dessin des residus contre les X dans PROC REG */
plot R.*X1='1' R.*X2='2' R.*X4='4' R.*X5='5' / vref = 0 ;
```

QQ-Plot (avec SAS/INSIHT)



Le tracé QQ-Plot montre un assez bon ajustement à la loi normale.

Test d'homoscédasticité et tracé du QQ-PLOT avec PROC REG.

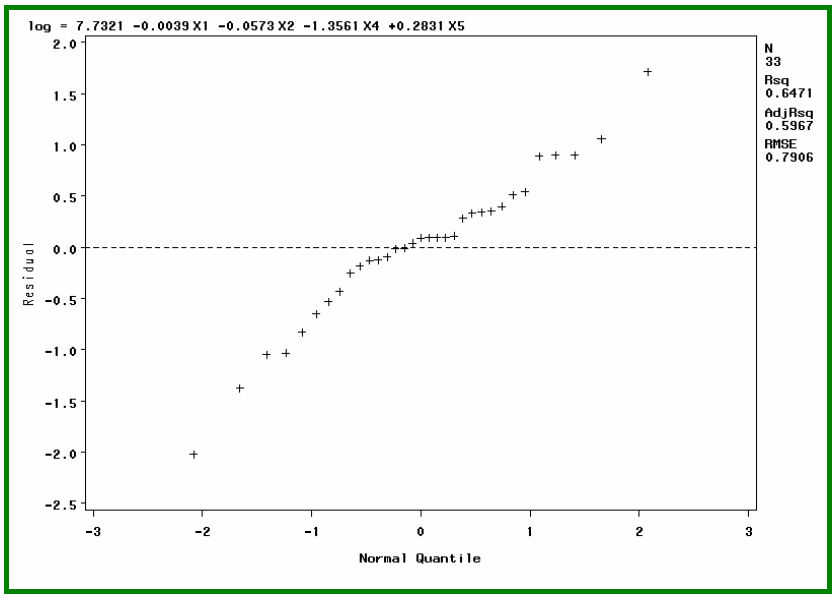
```
/*option SPEC + dessin QQ_Plot */
proc reg data=libreg.chenilles; title 'Modèle Y = X1 X2 X4 X5
';
LOG : model LOG=X1 X2 X4 X5 / SPEC ;
run ;
/* QQ Plot */
plot R.* NQQ. ;
quit;
```

Modèle Y = X1 X2 X4 X5

The REG Procedure
Model: LOG
Dependent Variable: log

Test d'indication du Premier et du Second		
DF	Khi 2	Pr > Khi 2
14	12.09	0.5991

L'homoscédasticité des résidus n'est pas rejetée.



4.3. Influence d'observations

Dans le but d'avoir une régression plus robuste, il faut détecter les observations influentes, détection qui commence là-aussi par des graphiques soit de Y contre les X_j, soit des résidus contre les X_j.
Dans la figure 4.3, e_i est nul car l'observation i est influente à cause de son caractère atypique.

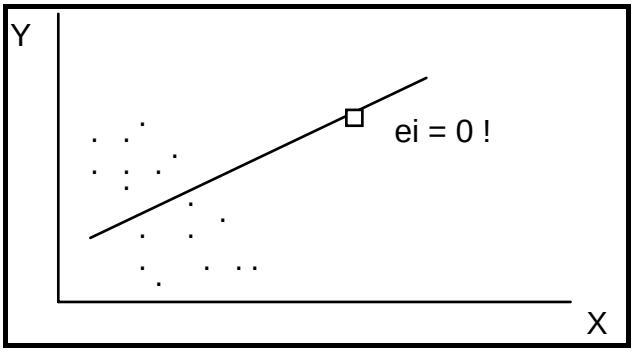


Figure 4.3 : Exemple d'observation à fort effet de levier

SAS calcule une série d'indicateurs par les 2 options R et INFLUENCE de l'instruction MODEL. L'ouvrage « Regression Diagnostics » de Belsley D.A., Kuh K. et Welsh R.E. (1980) en est la référence de base.

Ces indicateurs sont basés sur des détections de l'influence selon des mesures différentes, donc détecteront des influences de nature différente : on distinguera donc des observations influentes sur la régression, ou suspectes, ou atypiques, ce dernier terme étant plutôt recommandé.

Les mesures peuvent être classées en 3 groupes :

- détection d'un effet de levier de l'observation, donnant un résidu petit : *leverage* ;
- détection de résidu grand donc observation atypique ;
- détection d'un grand effet sur l'ajustement, ou les coefficients, ou la précision.

4.3.1. Hat matrice et leverages

On utilise ici le modèle sous sa forme matricielle :

$$Y = \begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \cdots \\ Y_n \end{vmatrix} \qquad X = \begin{vmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 \\ 1 \\ \cdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{vmatrix} \qquad \beta = \begin{vmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{vmatrix} \qquad \varepsilon = \begin{vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdots \\ \varepsilon_n \end{vmatrix}$$

L'estimation des moindres carrés est le vecteur $B = (X'X)^{-1}X'Y$.

D'où l'ajustement $\hat{Y} = X(X'X)^{-1}X'Y$

En notant $H = X(X'X)^{-1}X'$, on obtient $\hat{Y} = HY$ et $e = (I - H)Y$

Cette matrice est nommée H pour *Hat matrice* car hat se traduit par chapeau, et $\hat{Y}$ se dit *Y chapeau*.

Dans l'espace R^n, H est la matrice de la projection de Y sur l'espace engendré par les variables régresseurs X (espace de dimension (p+1)) : c'est donc la matrice d'un projecteur, dont deux propriétés sont : H' = H, et trace(H) = p+1.

H est une matrice carrée (n,n), dont la diagonale comporte les n coefficients h_{ii} .De l'expression matricielle de H, on déduit $h_{ii} = x_i'(X'X)^{-1}x_i$.

Les coefficients h_{ii} ne comportent donc que des données relatives aux variables explicatives Xj .

Les « *leverages* » (leviers) des observations sont ces n valeurs h_{ii}.

Un levier représente l'influence de l'observation i sur la valeur ajustée $\hat{Y}_i$, à cause des valeurs xi prises par les variables en i.

On peut montrer que $h_{ii} = (1/n) + (x_i - x_c)'(X_c'X_c)^{-1}(x_i - x_c)$, où ($x_i$ - x_c) est la différence entre le vecteur des valeurs des variables pour l'observation i, et le vecteur des valeurs moyennes, et X_c la matrice de taille (n,p) des valeurs centrées.

Le levier en i est donc une « distance » entre les valeurs des X prises en i et les valeurs moyennes calculées sur les n observations

Des différentes propriétés de H, on déduit:

$$h_{ii} = \sum_{j=1,n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

$$\text{trace}(H) = p + 1 \Rightarrow \sum_{i=1}^{p} h_{ii} = p + 1$$

$$\Rightarrow 1/n \leq h_{ii} \leq 1$$

et aussi des formules concernant les variances : var($\hat{Y}_i$) = $\sigma^2 h_{ii}$ et var(ei) = $\sigma^2(1-h_{ii})$.

On en conclut que h_{ii} est toujours plus petit ou égal à 1.

Règle (colonne Hat Diag H): Si les leviers étaient tous égaux, la valeur commune serait (p+1)/n. De façon empirique, un levier supérieur à 2(p+1)/n est suspect.

4.3.2. Résidus studentisés internes

Ils sont appelés en anglais *Standardized Residuals* ou STUDENT.

On connaît la variance de chaque résidu : $\mathrm{var}(e_i) = \sigma^2(1-h_{ii})$. Dans cette formule, on estime σ^2 par $s^2 = \dfrac{\sum_i e_i^2}{n-(p+1)}$ = MSE; donc le résidu standardisé est :

$$\boxed{r_i = e_i \Big/ s\sqrt{1-h_{ii}}}\ .$$

Bien que numérateur et dénominateur ne soient pas indépendants, on considère que cette quantité r_i suit une loi de Student à (n-1-(p+1)) = (n-p-2) ddl, d'où le nom STUDENT.

Règle (colonne Student) : r_i sera suspect si |ri| > 2 (quantile de la loi de Student (1-α/2), pour le seuil α=5%, avec l'approximation par une loi normale si n grand).

4.3.3. Résidus studentisés externes

Ils sont appelés en anglais *Studentized Residuals* ou RSTUDENT.

On remplace dans l'expression de r_i l'estimation s par $s_{(-i)}$ qui est l'estimation de s obtenue en refaisant l'ajustement du modèle sans l'observation i, ce qui rend e_i indépendant de $s_{(-i)}$: ceci donne

$$\boxed{r_{(-i)} = e_i \Big/ s_{(-i)}\sqrt{1-h_{ii}}}$$

Règle (colonne RStudent) : $r_{(-i)}$ suit aussi une loi de Student à (n-1-(p+1)) ddl et sera donc aussi suspecte si |r(-i)| > 2. Certains auteurs préconisent d'autres quantiles à un seuil fixé α pour $r_{(-i)}$: quantile 1-α/2n plutôt que 1-α/2 .

4.3.4. Mesure globale de l'influence sur le vecteur des coefficients: Distance de COOK

Pour chaque observation i, on calcule une « distance » entre le vecteur B des coefficients de la régression et le vecteur B(-i) obtenu en refaisant la régression sans l'observation i : la distance se mesure à l'aide de (X'X) et est normée par s^2, estimation de σ^2.

$$\boxed{(\mathrm{COOKD})_i = \frac{|B-B(-i)|'(X'X)|B-B(-i)|}{(p+1)s^2} = \frac{r_i^2 h_{ii}}{(p+1)(1-h_{ii})}}$$

Règle : (colonne Cook's D) : La distance de COOK étant normée, une valeur supérieure à 1 est suspecte [22].

[22] Certains auteurs suggèrent une limite de 4/(n-p-1), la calibration à 1 pouvant laisser passer des valeurs influentes

4.3.5. Influence sur chacun des coefficients : DFBETAS

Pour chaque variable, on calcule la différence entre le coefficient estimé bj et celui obtenu avec l'estimation sans l'observation i, bj(-i).
Avec standardisation, on obtient pour chaque variable explicative j :

$$\boxed{(\mathrm{DFBETAS})_i^j = \frac{(b_j - b_j(-i))}{s(-i)\sqrt{(X'X)_{j,j}^{-1}}}}$$

Règle : Empiriquement, un DFBETAS dont la valeur absolue est plus grande que $2/\sqrt{n}$ est suspect.

Utilisation conjointe COOKD et DFBETAS : S'il y a beaucoup de variables, on regarde d'abord les observations globalement influentes (COOKD élevé), puis pour cette observation quelle(s) variable(s) cause(nt) cette influence (DFBETAS).

4.3.6. Précision des estimateurs : COVRATIO

La quantité Mean Square Error (MSE) mesure la précision globale de l'estimation : MSE petit indique une bonne précision. MSE est aussi la variance des résidus.

Ici, on mesure la précision en utilisant une variance « généralisée », évaluée par : $s^2 \| (X'X)^{-1} \|$ calculée avec et sans l'observation i ((la notation || (..)|| désigne le déterminant de la matrice) :

$$\boxed{(\mathrm{COVRATIO})_i = \frac{s^2(-i)\left\|(X'X)_{(-i)}^{-1}\right\|}{s^2\left\|(X'X)^{-1}\right\|}}$$

Donc $(\mathrm{COVRATIO})_i$ plus grand que 1 indique que le fait de mettre l'observation i augmente la précision, alors qu'une valeur plus petite que 1 indique une diminution de la précision.

Règle : Belsley, Kuh et Welsh suggèrent qu'un écart à l'unité dépassant 3(p+1)/n est grand.

4.3.7. Influence sur la valeur ajustée: DFFITS

Pour chaque observation i, $(\mathrm{DFFITS})_i$ donne la différence entre la valeur ajustée pour l'observation i et la valeur prédite de Y pour i dans le modèle estimé sans cette observation i. Un grand écart indiquera une forte modification dans la valeur ajustée par le modèle quand l'observation i est retirée.
Avec une standardisation à s(-i):

$$\boxed{(\mathrm{DFFITS})_i = \frac{\hat{Y}_i - \hat{Y}_i(-i)}{s(-i)\sqrt{h_{ii}}} = r(-i)\sqrt{\frac{h_{ii}}{1-h_{ii}}}}$$

Règle : DFFITS est déclaré suspect s'il est en valeur absolue plus grand que $2\sqrt{(p+1)/n}$.

4.3.8. Coefficient global PRESS

Predicted Residuals Sum of Squares = $\boxed{\text{PRESS}= \sum_{i=1,n}\left(Y_i - \hat{Y}_i(-i)\right)^2}$

Ce coefficient (unique) est calculé en faisant n estimations $\hat{Y}_i(-i)$ obtenues en enlevant une observation. Il devrait donc être égal à la somme des carrés des résidus du modèle avec toutes les observations (SSResidus) si aucune observation ne pose problème.

Des coefficients PRESS individuels $\left(Y_i - \hat{Y}_i(-i)\right)^2$ peuvent être obtenus uniquement dans la table SAS créé par l'instruction OUTPUT. Ces coefficients peuvent être comparés aux coefficients DFFITS, comme le montre la figure 4.4 ci-dessous :

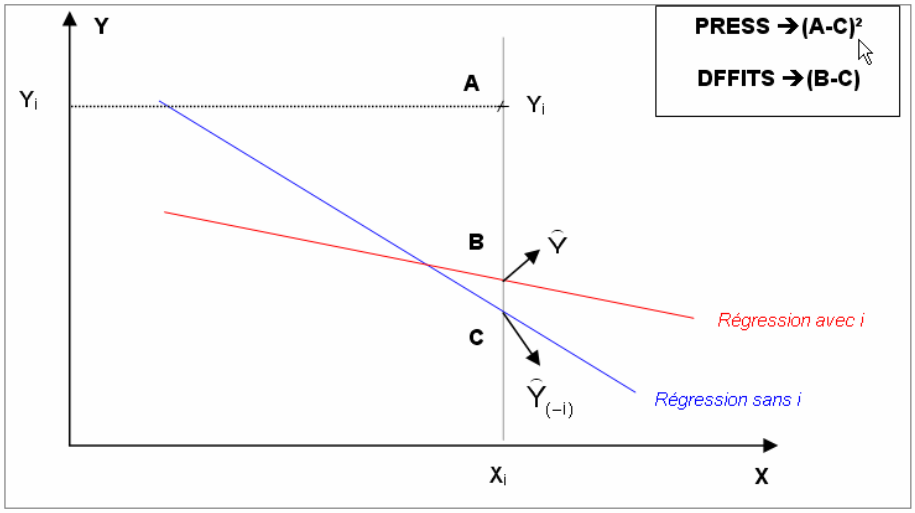


Figure 4.4 : Illustration de PRESS et DFFITS

4.3.9. Comment obtenir les mesures d'influence dans SAS

Dans PROC REG

Elles sont affichées en sortie par des options de l'instruction MODEL.

- Option R → pour toutes les observations, le résidu et son écart-type, les résidus standardisés STUDENT, la distance de COOK et un dessin indiquant la position du résidu par rapport à l'intervalle [-2 ; +2] ;
- Option INFLUENCE → pour toutes les observations, levier, résidu studentisé externe RSTUDENT, COVRATIO, DFFITS et DFBETAS de chaque coefficient.

Certaines mesures peuvent être stockées dans une table de sortie SAS (table possédant n lignes au moins), à l'aide de l'instruction OUTPUT : on crée des variables en utilisant les mots-clefs R, STUDENT, RSTUDENT, H, COOKD, DFFITS et PRESS.
L'option PRESS de l'instruction MODEL permet d'obtenir le coefficient PRESS « global », qui sera affiché en sortie, et stocké dans la table de l'instruction OUTPUT si cette instruction existe.

Dans SAS/INSIGHT

Une fois que l'on a exécuté le modèle, on peut ajouter à la table SAS sur laquelle on travaille, des variables à l'aide du menu Vars. Les variables ajoutées ont alors le nom indiqué entre parenthèses ci-dessous (où Y est le nom de la variable réponse du modèle) :

- Hat Diag, (H_Y)
- Residual (R_Y)
- Standardized residual (RS_Y)
- Studentized residual (RT_Y)
- Cook's D (D_Y)
- Dffits (F_Y)
- Covratio (C_Y)
- Dfbetas (BY_Intercept, BY_X1, BY_X2, etc.)

4.3.10. Tableau récapitulatif

	Std Err Residual	**Student Residual**	**Rstudent**	**Hat Diag H**
signifiant	estimateur de l'erreur-type du résidu i	résidus studentisés internes, appelés standardized residual dans SAS-Insight	résidus studentisés externes, appelés studentized residual dans SAS- Insight	levier de l'obs. i
objet	permet de calculer l'intervalle de confiance autour du résidu i	test de significativité du résidu i	à comparer avec Student Residual écart-type calculé en retirant l'obs. i	mesure l'influence de l'obs.i à cause des valeurs xi
valeurs critiques		2	2	$\frac{2(p+1)}{n}$
Règle de décision		$\lvert \text{Student residual} \rvert > 2$ alors le résidu i est significativement $\neq 0$	$\lvert \text{RStudent} \rvert > 2$ alors l'observation i nécessite une investigation !	$h_{ii} > \frac{2(p+1)}{n}$ nécessite une investigation
Option de PROC REG	***R***	***R***	***Influence***	***Influence***

	Cook's D	Df betas	Cov Ratio	Dffits
signifiant	distance de Cook	DFBETAS relatif à chaque coefficient β_j	Ratio de MSE sans et avec l'observation i	statistique DFFITS
objet	mesure le changement en retirant l'obs. i, sur les estimations de l'ensemble des coefficients	mesure normalisée de l'effet de l'obs. i sur l'estimation, pour chaque coefficient β_j	mesure l'effet de l'obs. i sur la précision	mesure normalisée du changement dans la valeur prédite, avec et sans l'obs. i
valeurs critiques	1 ou $\frac{4}{(n-p-1)}$	$\frac{2}{\sqrt{n}}$	$\frac{3(p+1)}{n}$	$2\sqrt{\frac{(p+1)}{n}}$
Règle de décision	$\mathrm{CookD} > 1$ alors l'observation i est influente globalement	$\lvert \mathrm{Dfbetas} \rvert > \frac{2}{\sqrt{n}}$ indique une influence de l'obs. i sur l'estimation de β_j	$\lvert \mathrm{Covratio} - 1 \rvert > \frac{3(p+1)}{n}$ nécessite une investigation	$\lvert \mathrm{Dffits} \rvert > 2\sqrt{\frac{(p+1)}{n}}$ indique une influence de l'obs. i sur $\hat{Y}_i$
Option de PROC REG	***R, Influence***	***Influence***	***R***	***R***

On trouvera dans le programme SAS ci-dessous une macro CRITIQUE
(avec comme paramètres : n nombre d'observations, p nombre de régresseurs, et b0 indicateur de la présence d'une constante)
permettant d'afficher les valeurs critiques des différentes mesures d'influence.

```
%MACRO CRITIQUE     (n=  ,          /* nombre d'observations*/
                     p=  ,          /* nombre de régresseur */
                     b0=1           /* bo=1 si constante
(intercept) */
                 );
data seuil;
n=&n; p=&p; b0=&b0;
dcook=4/(&n-&p-&b0);
hat_diag=2*(&p+&b0) / &n;
covratio=3*(&p+&b0) / &n;
dffits=2*sqrt((&p+&b0) / &n);
dfbetas=2/sqrt(&n);
proc print data=seuil;
%mend;
/*_____________________________________________________________*/
*exemple  appel de la macro ;
%critique(n= 44 ,p= 4 ,b0=1);
run;
```

4.3.11. Exemple

On utilise les données « Processionnaire du pin » issu du livre de Tomassone et *al.*(1983) et on se limite de nouveau au modèle Y = log = f(X1 X2 X4 X5).

```
/* appel de la macro */
Title 'Valeurs critiques pour n = 33 et p = 4';
%critique(n= 33 ,p= 4 ,b0=1);
run;

title ' influence des observations ';
proc reg data=libreg.chenilles;;
LOG : model LOG=X1 X2 X4 X5 /R influence ;
run;
/* exemple de stockage des criteres */
output out=influence H=levier COOKD=dcook STUDENT = rsi
RSTUDENT = rse ;
quit;
```

Voici les valeurs limites pour les coefficients d'influence :

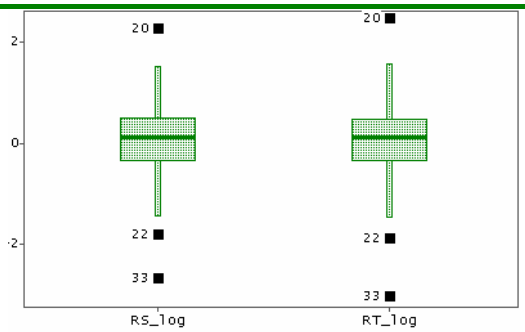
Obs	n	p	b0	dcook	hat_diag	covratio	dffits	dfbetas
1	44	4	1	0.10256	0.22727	0.34091	0.67420	0.30151

La table des valeurs des coefficients d'influence de toutes les observations est donnée ci-après. Mais plutôt que de rechercher les valeurs limites dans cette table, de simples Box-Plot permettent de les repérer rapidement.

Obs	R_log	RS_log	RT_log	H_log	D_log	Blo_Intercept	Blo_X1	Blo_X2	Blo_X4	Blo_X5	F_log	C_log
1	0.33599	0.44584	0.43937	0.09147	0.00400	0.09843	-0.05828	-0.06801	-0.04923	0.05234	0.13941	1.27397
2	0.39625	0.53225	0.52532	0.11337	0.00724	0.00264	0.02666	-0.01167	-0.14634	0.15830	0.18784	1.28572
3	-0.12740	-0.17506	-0.17200	0.15276	0.00111	-0.02891	0.00102	-0.00445	0.03108	-0.00407	-0.07303	1.40794
4	0.09551	0.12696	0.12471	0.09480	0.00034	0.01497	-0.00089	0.00104	-0.00649	-0.00786	0.04036	1.32122
5	-0.01087	-0.01442	-0.01416	0.09104	0.00000	0.00078	-0.00140	-0.00116	-0.00111	0.00244	-0.00448	1.31951
6	0.89723	1.15893	1.16637	0.04120	0.01154	0.12929	-0.10515	-0.04965	0.04266	-0.03267	0.24177	0.97836
7	0.54068	0.77260	0.76690	0.21656	0.03300	-0.13312	0.18854	0.17878	-0.00538	-0.13641	0.40321	1.37459
8	-0.43550	-0.63061	-0.62369	0.23707	0.02471	0.06998	0.08085	-0.28038	-0.07361	0.01658	-0.34766	1.46360
9	0.34463	0.46397	0.45737	0.11742	0.00573	0.14429	-0.10171	-0.03953	-0.05180	0.03991	0.16683	1.30755
10	1.06078	1.53641	1.57665	0.23745	0.14701	0.48930	-0.63930	0.26565	0.47664	-0.47132	0.87979	1.01264
11	-0.01681	-0.02399	-0.02355	0.21461	0.00003	-0.00545	0.00714	0.00348	-0.00904	0.00654	-0.01231	1.52700
12	0.10911	0.16228	0.15943	0.27685	0.00202	0.02039	-0.04996	0.05580	-0.02903	0.05115	0.09865	1.65083
13	0.09232	0.12927	0.12697	0.18399	0.00075	0.04025	-0.01407	-0.04281	-0.00605	-0.00224	0.06029	1.46547
14	0.51470	0.69675	0.69020	0.12704	0.01413	0.10261	-0.07666	-0.14632	-0.05782	0.12497	0.26330	1.25892
15	0.28053	0.37025	0.36447	0.08165	0.00244	0.04857	-0.02676	-0.02181	-0.06915	0.07560	0.10868	1.27440
16	0.35179	0.49872	0.49193	0.20406	0.01275	-0.18643	0.12665	0.09137	0.12233	-0.10837	0.24908	1.44116
17	-0.25012	-0.34526	-0.33976	0.16046	0.00456	0.10308	-0.10651	0.03639	0.00433	-0.02423	-0.14854	1.39851
18	-0.18215	-0.24595	-0.24178	0.12257	0.00169	-0.01182	-0.00607	-0.01127	0.07785	-0.07214	-0.09036	1.35226
19	-0.53336	-0.72135	-0.71502	0.12543	0.01493	-0.00664	-0.05219	-0.11461	0.17012	-0.09230	-0.27079	1.24866
20	1.71056	2.29211	2.49725	0.10908	0.12864	-0.49654	0.35204	0.60098	0.06252	-0.13427	0.87379	0.47631
21	0.89615	1.29548	1.31207	0.23453	0.10284	-0.38213	0.53134	-0.39635	-0.08150	0.12191	0.72625	1.15033
22	-1.37360	-1.79024	-1.86815	0.05826	0.03965	0.11629	-0.26821	0.15040	0.14571	-0.09153	-0.46464	0.69354
23	0.88622	1.16287	1.17053	0.07090	0.02064	-0.07800	0.05356	-0.02543	-0.07646	0.16123	0.32336	1.00791
24	-0.82583	-1.15686	-1.16418	0.18483	0.06069	0.36554	-0.44270	0.21731	-0.04419	0.02911	-0.55435	1.15178
25	0.09451	0.13298	0.13063	0.19199	0.00084	-0.02751	0.01017	0.05453	0.00887	-0.01224	0.06368	1.47973
26	0.09002	0.11919	0.11707	0.08759	0.00027	-0.00737	0.01298	-0.00722	-0.02162	0.02507	0.03627	1.31123
27	0.03982	0.05739	0.05636	0.22964	0.00020	-0.00827	0.00422	-0.01313	0.02032	-0.01208	0.03077	1.55604
28	-0.65248	-0.93937	-0.93733	0.22821	0.05219	-0.01506	-0.16289	0.10754	-0.11565	0.29823	-0.50970	1.32416
29	-0.09597	-0.12660	-0.12435	0.08079	0.00028	-0.02556	0.01179	0.01527	0.00357	0.00298	-0.03687	1.30110
30	-1.05040	-1.39433	-1.41937	0.09215	0.03947	-0.24601	0.20872	0.00778	-0.21675	0.27484	-0.45222	0.92192
31	-1.03359	-1.43776	-1.46704	0.17328	0.08666	-0.18417	0.28956	-0.07650	0.26750	-0.46377	-0.67164	0.98872
32	-0.12787	-0.19257	-0.18922	0.29470	0.00310	-0.03068	0.04274	0.04959	-0.10197	0.07997	-0.12231	1.68934
33	-2.02086	-2.65655	-3.01635	0.07429	0.11327	0.09749	0.20004	-0.49187	-0.22716	0.06711	-0.85449	0.30330

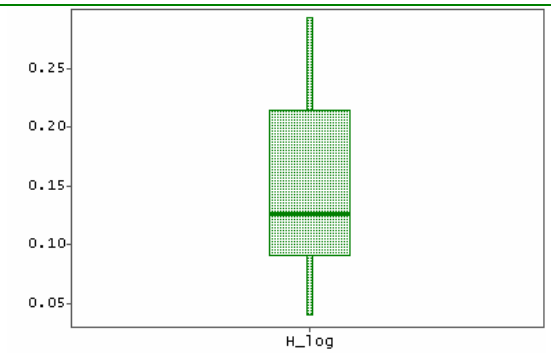
Dessins BOX-PLOT des coefficients d'influence :
Ce type de représentation est très riche en information sur la forme d'une variable, et l'existence de valeurs différentes des autres (outliers). On en trouvera l'explication détaillée dans Le Guen (2001) [23].

[23] http://matisse.univ-paris1.fr/leguen/leguen2001b.pdf

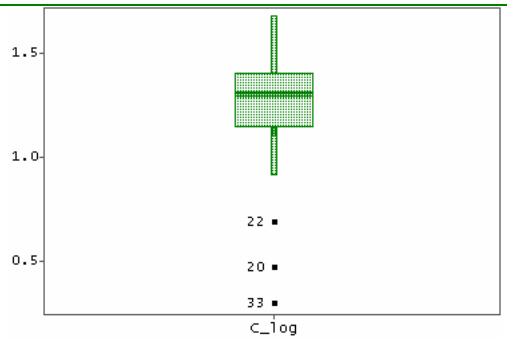


Résidus standardisés

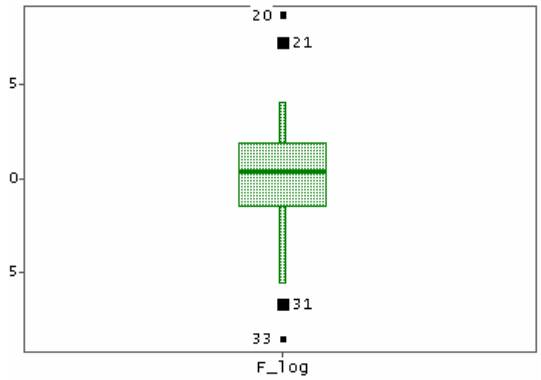
Internes **Externes**



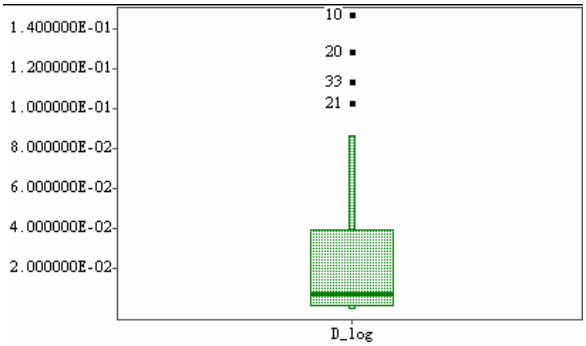
Leviers



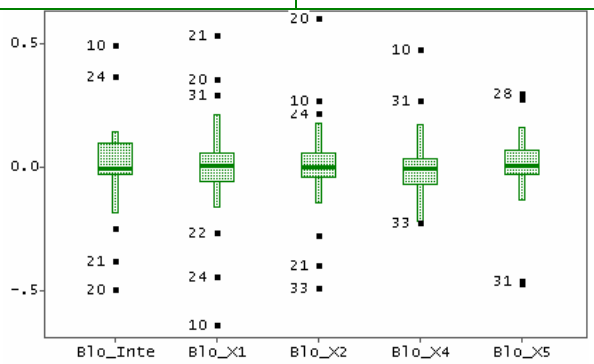
CovRatio



Dffits



DCook



Dfbetas

La distribution des leviers ne révèle pas d'outliers, mais au vu des valeurs, on constate que quelques observations ont cependant un levier un peu trop grand (> 0.23) : 8, 10, 12, 21 et 32.
Les observations 20 et 33 ont des résidus « grands » (|STUDENT| et |RDSTUDENT| > 2) et donc un grand effet sur la précision (COVRATIO -1) > 0.34 car quand on les enlève de la régression, SSE diminue fortement.
Et elles influent également sur |DFFITS| >> 0.67. Les observations 10 et 21 ont une forte incidence sur l'ensemble des coefficients (DCOOK > 0.10) et donc sur certains coefficients (DFBETAS > 0.30), ceux de X1 et X2 pour les 2 observations, celui de X4 pour l'observation 10.

Pour résumer, les observations 10 20 21 et 33 seraient atypiques. Cependant, au vu des données, où ces observations sont des placettes issues d'un plan d'expérience, on peut penser qu'il n'est pas judicieux de les retirer de la régression.

Le dessin humoristique de la figure 4 .5 ci-dessous (publié dans le manuel du logiciel OSIRIS, mais l'auteur nous est inconnu) illustre le mauvais réflexe que pourrait avoir le statisticien au vu de données atypiques !

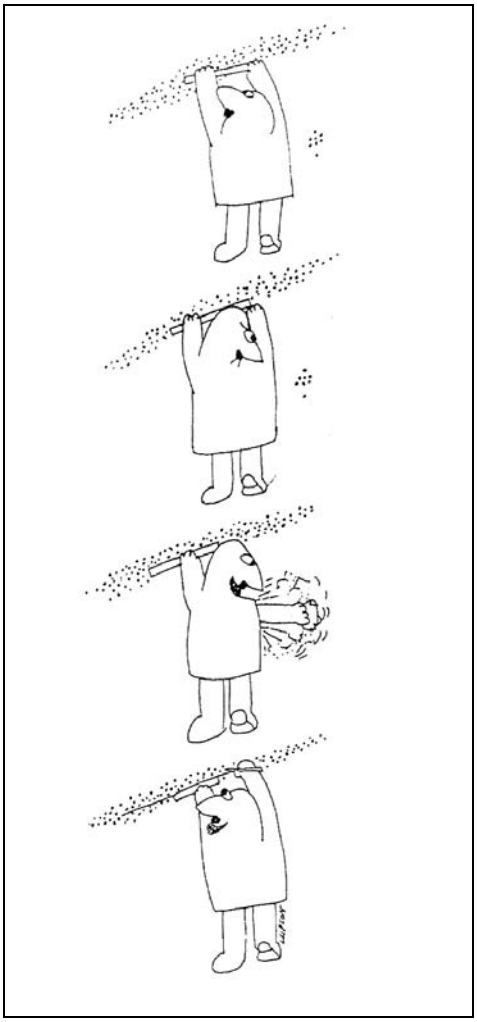


Figure 4.5 : Que faire en présence de données atypiques ?

4.4. Colinéarité des régresseurs

Différents symptômes sont révélateurs de problèmes de colinéarité :

- De grandes corrélations entre les régresseurs ;
- Un grand changement dans les coefficients quand on ajoute ou enlève un régresseur ;
- Des coefficients non significatifs alors que le test global d'analyse de variance sur tous les coefficients est significatif ;
- La non significativité et/ou une très grande variance pour le coefficient d'un régresseur théoriquement important dans le modèle ;
- Un coefficient de signe opposé à celui auquel on s'attendait.

Ce sont des problèmes d'inversion de X'X, qui entrainent une augmentation des variances des coefficients, et donc leur instabilité car $\mathrm{Var}(\beta) = \sigma^2 (X'X)^{-1}$. Et la non-inversion de X'X se rencontre quand il existe des combinaisons linéaires entre les colonnes de X.

L'exemple simple de la figure 4.6 illustre le phénomène d'instabilité dû à la liaison entre deux régresseurs.

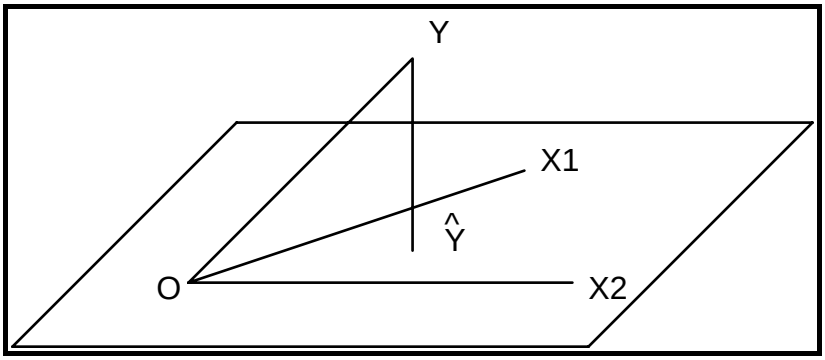


Figure 4.6 : Projection de Y dans l'espace de 2 régresseurs corrélés

Dans l'espace R^n, les cosinus d'angle entre variables sont les corrélations (voir chapitre 1, §1.2.5): Ici X1 et X2 étant très corrélées, l'angle entre les 2 vecteurs est petit, ce qui rend le sous-espace (X1,X2), et donc la projection, instables[24].

4.4.1. Méthodes basées sur l'étude de la matrice X'X

La matrice X possède n lignes et (p+1) colonnes :

$$X = (x_{ij}) = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{1p} \\ 1 & x_{21} & x_{22} & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{np} \end{bmatrix}$$

La matrice (X'X) est une matrice carrée (p+1), symétrique, dont les éléments sont calculés ainsi :

$$\boxed{\begin{aligned} (X'X)_{11} &= n \\ (X'X)_{kk} &= \sum_{i=1,n} x_{ik}^2, \forall k \neq 1 \\ (X'X)_{kh} &= \sum_{i=1,n} x_{ik} x_{ih}, \forall k,h \neq 1 \\ (X'X)_{k1} &= \sum_{i=1,n} x_{ik} = n.\text{moyenne}(X_k), \forall k \neq 1 \end{aligned}}$$

En étudiant les leviers au §4.3.1, on a utilisé les variables régresseurs centrées. Si on considère la matrice X_c de taille (n,p) ayant en colonnes les p régresseurs centrés, alors $(X_C'X_C) = n$. COV où COV est la matrice de variance-covariance des variables X_j. Quand les variables sont non-corrélées, alors $(X'_C X_C)$ est une matrice diagonale puisque toutes les covariances sont nulles.

[24] Le site de Chong Ho Yu illustre de façon très amusante ce problème d'instabilité http://creative-wisdom.com/computer/sas/collinear_subject_space.html

Donc X'X est égal à n.R si on travaille sur des régresseurs centrés et réduits, où R est la matrice carrée symétrique des coefficients de corrélation entre les p régresseurs.

Etude de la matrice de corrélation des régresseurs

Cette étude se révèle intéressante.

- Si les régresseurs ne sont pas corrélés entre eux, cette matrice, notée R, n'a que des 1 sur sa diagonale, et 0 ailleurs : c'est la matrice identité.

- La matrice R est symétrique définie et positive, et de rang p. Les valeurs propres de R sont donc en nombre de p, elles sont positives et leur somme vaut p (trace de R). Quand il n'y a aucune liaison entre les régresseurs, elles sont toutes égales à 1 car R est la matrice identité. Sinon, on aura des valeurs propres plus petites et mêmes proches de 0 : l'examen des valeurs propres révèlera donc les problèmes de liaison entre les régresseurs.

- $R^{(-1)}$, matrice inverse de R, est également riche en informations. En effet on peut montrer que ses éléments diagonaux $\left(R_j^{(-1)}\right)$ sont égaux à $1/(1 - R^2_j)$ où R^2_j est le coefficient de corrélation multiple de la régression avec constante de X_j sur les (p-1) autres variables.

- Si on définit un modèle avec les régresseurs centrés et réduits, en remplaçant l'expression de (X'X) dans l'estimateur des moindres carrés, on montre (cf. Woolridge (2000)) que le vecteur des coefficients B_c de ce modèle est $B_c = \frac{1}{n} R^{-1} X'Y$, et la matrice de variance-covariance vaut $\mathrm{Var}(B_c) = \frac{\sigma^2}{n} R^{-1}$.

Une étude préalable de la matrice de corrélation des régresseurs, complétée éventuellement par une Analyse en Composantes Principales, s'impose donc et permet de visualiser les liaisons entre variables.
REG possède des options TOL, VIF et COLLIN, COLLINOINT pour détecter des problèmes de colinéarité selon deux optiques différentes.
SAS/ INSIGHT donne les indices TOL et VIF par défaut, et affiche la table COLLIN uniquement.

4.4.2. Variance Inflation Factor

On a vu au §4.4.1 que la matrice de variance –covariance des coefficients du modèle où les régresseurs sont centrés et réduits est R^{-1}, à un facteur près. Donc l'élément diagonal j de cette matrice mesure comment la variance du coefficient de X_j sera augmentée par la colinéarité.

Pour chaque variable X_j , on nomme cet élément VIF_j (*Variance Inflation Factor* ou *inflation de variance*) : $VIF_j = 1/(1 - R^2_j)$ où R^2_j est le coefficient de corrélation multiple de la régression avec constante de X_j sur les (p-1) autres variables.

S'il y a colinéarité, alors R^2_j est proche de 1, donc VIF_j est grand. Comme la loi de ce coefficient n'est pas connue, Belsley et *al.* ont défini un seuil limite de façon empirique.

Règle : une valeur de VIF plus grande que 10 révèle un problème***.***

Tomassone (1983) propose de calculer un indice global de colinéarité défini comme la somme des VIF de tous les régresseurs : $\boxed{I = \frac{1}{p}\left(\sum_{j=1}^{p} VIF_j\right)}$

Remarque*:* la tolérance (option TOL) est définie comme l'inverse de la variance inflation
TOLj = 1/VIFj

4.4.3. Condition index et variance proportion

On a signalé au §4.4.1 que l'étude les valeurs propres de la matrice de corrélations révèle les problèmes de liaisons entre les régresseurs.

Cette étude se fait aussi par l'analyse en composantes principales (ACP), qui consiste à transformer des variables pour obtenir d'autres variables orthogonales, qui sont des combinaisons linéaires des premières, appelées composantes principales (cf. Saporta (2006) ou Tenenhaus (1994)).

Plus précisément si Z est la matrice (n,p) des variables initiales centrées et réduites , on construit la matrice W (n,p) des variables orthogonales avec la relation W = Z U. En ACP, on démontre que la matrice U est la matrice des vecteurs propres normés de R, associés aux p valeurs propres $(\lambda_k, k = 1,2,...,p)$, qui sont positives car R est symétrique définie positive. On les ordonne de la plus grande à la plus petite : une liaison parfaite entre les variables Z entraine une nullité des dernières valeurs propres.
On montre également que la variance d'une composante W_k est égale à λ_k.

Si on construit le modèle avec les variables W [25], on trouve donc la solution des moindres carrés $Y = W\hat{c} + e$, avec $\hat{c} = (W'W)^{-1}W'Y$.
Et la matrice de variance-covariance des coefficients est $Var(\hat{c}) = \sigma^2(W'W)^{-1}$; pour un coefficient : $var(\hat{c}_k) = \frac{\sigma^2}{n\lambda_k}$.
Comme les variables sont orthogonales et de variance égale à la valeur propre, on en déduit $Y = W\hat{c} + e = \Omega^{-1}W'Y$, avec $\Omega^{-1} = Diag\left[\left(\frac{1}{n\lambda_k}\right), k = 1,..,p\right]$ matrice diagonale.

[25] On trouvera dans le chapitre 6 du livre Tomassone et *al* (1992), le principe de cette méthode qu'il appelle « régression orthogonalisée »

On passe facilement de $\hat{c}$ à $\hat{\alpha}$ car $Y = W\hat{c} + e = ZU\hat{c} + e = Z\hat{\alpha} + e$, et donc $\hat{\alpha} = U\hat{c}$. La matrice de variance-covariance est $Var(\hat{\alpha}) = UVar(\hat{c})U'$; pour le coefficient du régresseur j, $var(\hat{\alpha}_j) = \frac{\sigma^2}{n}\sum_{k=1}^{p}\frac{U_{jk}^2}{\lambda_k}$: des valeurs propres faibles entrainent donc de grandes variances des coefficients.

Les indices de colinéarité

A - Tout d'abord, l'édition des valeurs propres donnera des informations sur l'existence de colinéarité.

De façon générale, on calcule les valeurs propres de la matrice (X'X) du modèle, préalablement transformée pour avoir uniquement la valeur 1 sur les éléments diagonaux. Il y a donc (p+1) valeurs propres : c'est l'option COLLIN.
Si on travaille sur un modèle avec les régresseurs centrés comme ci-dessus, il y aura p valeurs propres : c'est l'option COLLINOINT.

On édite ces valeurs propres de la plus grande λ_1 à la plus petite λ_L, (L=p ou p+1). Une valeur propre nulle révèle l'existence d'une dépendance linéaire entre les colonnes de X, donc une colinéarité.

On nomme « Condition Index » le rapport $\boxed{CI = \sqrt{\frac{\lambda_1}{\lambda_k}}}$, appelé aussi « indice de conditionnement ». Le dernier de ces rapports $CI = \sqrt{\frac{\lambda_1}{\lambda_L}}$ (L=p ou p+1) est nommé « Condition Number ».

Comme la loi de ce coefficient n'est pas connue, Belsley et *al.* ont défini un seuil limite de façon empirique :

Règle (colonne Condition Index) : une valeur grande met en évidence un problème ; empiriquement CI > 30 avec l'option COLLINOINT, ou CI > 100 avec COLLIN.

Remarque : On peut définir une « indice de multicolinéarité » en calculant la moyenne des inverses des valeurs propres (cf. Foucart 2006, 2007) : Cet indice serait calculé comme $I = \frac{1}{p}\left(\sum_k \frac{1}{\lambda_k}\right)$ si on considère que les régresseurs sont centrés et réduits (p valeurs propres).

B – Ensuite, pour chaque valeur propre et donc chaque CI, sont données des « *VARiance PROPortions* », qui indiquent quelles variables sont responsables de la colinéarité révélée par cette valeur propre.
En effet, on a vu que la matrice de variance-covariance des coefficients $\hat{\alpha}$ de la régression sur les variables centrées et réduites est $Var(\hat{\alpha}) = UVar(\hat{c})U'$ et que pour un coefficient j, $var(\hat{\alpha}_j) = \frac{\sigma^2}{n}\sum_{k=1}^{p}\frac{U_{jk}^2}{\lambda_k}$.

La colonne « proportion de variance » pour le coefficient d'une variable j est le vecteur $\boxed{\text{Var.Prop.} = \left(\frac{U_{jk}^2}{\lambda_k}, k = 1,..,p \text{ ou } (p+1)\right)}$, normé pour que la somme de ses composantes soit égale à 1.

Règle : d'après Belsley, Kuh et Welsh, si les proportions de variance de plusieurs variables sont plus grandes que 0.50 pour un « condition index » grand, les variables correspondantes ont un problème de colinéarité entre elles.

Remarque *:* l'option COLLINOINT exclut la constante des estimateurs de coefficients; les p variables sont centrées et réduites et (X'X) est donc, à un coefficient près, la matrice de corrélation entre les p variables explicatives.
L'option COLLIN inclut la constante dans les estimations de coefficients. X contient donc la variable constante égale à 1. La matrice X'X, de taille (p+1), est normée pour avoir 1 sur la diagonale, mais les variables ne sont pas centrées.
Belsley, Kuh et Welsh recommandent de n'utiliser l'option COLLIN que si la constante a une interprétation physique. Centrer les variables (option COLLINOINT) consiste à supposer que la constante n'a pas d'effet sur la colinéarité des autres variables régresseurs. De plus, ceci est cohérent avec les calculs du « Variance Inflation Factor ». On trouvera dans l'article d'Hélène Rousse-Erkel (1990) des précisions et des prolongements aux travaux de Belsley et *al.* sur la colinéarité.

4.4.4. Remèdes en cas de multi-colinéarité

- Retirer certains régresseurs, principaux « responsables » de la colinéarité ;
- Les transformer par des ratios si on identifie le facteur commun de liaison ;
- Augmenter la taille n de l'échantillon avec le recueil d'autres observations ;
- Sélectionner les régresseurs, si p est trop grand par rapport à n ;
- Faire une Ridge-Regression [26] (transformer (X'X) en (X'X + kI)) (Hoerl et Kennard (1970)) ;
- Travailler sur les composantes principales issues des régresseurs [27] ;
- Faire une régression PLS ;
- Utiliser les méthodes de type « LASSO » de Tibshirani (1996) ;
- Etc. .

La régression RIDGE et la régression sur composantes principales peuvent être réalisées à l'aide des options RIDGE et PCOMIT de l'instruction Proc REG (voir l'annexe 1 pour la syntaxe).

Remarques :
- Quelques unes de ces méthodes sont décrites dans les articles de la R.S.A. de P.Cazes (1975) ou de R. Palm et A.F.Iemma (1995).
- La régression PLS (*Partial Least Square*) semble une méthode plus efficace que la régression Ridge ou la régression sur composantes principales en cas de

[26] Voir l'exemple au §4.4.5.
[27] C'est la « régression orthogonalisée » de Tomassone (§4.4.3.)

colinéarité, et s'applique aussi au cas où p est très grand par rapport à n (voir les publications de Tenenhaus (1995,1998)).

PROC ORTHOREG de SAS propose d'autres solutions pour réaliser une régression sur données « mal conditionnées », c'est à dire en cas de colinéarités des variables.

4.4.5. Exemple

On utilise les données « Processionnaire du pin » issu du livre de Tomassone et *al.*(1983), avec le modèle Y = log = f(X1 X2 X4 X5).

```
proc reg data=libreg.chenilles ; title 'option TOL VIF';
LOG : model log=X1 X2 X4 X5 /tol vif collinoint;
run;
quit;
```

Résultats estimés des paramètres							
Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|	Tolérance	Inflation de variance
Intercept	1	7.73214	1.48858	5.19	<.0001	.	0
X1	1	-0.00392	0.00115	-3.42	0.0019	0.89075	1.12265
X2	1	-0.05734	0.01939	-2.96	0.0062	0.97425	1.02643
X4	1	-1.35614	0.31983	-4.24	0.0002	0.17630	5.67209
X5	1	0.28306	0.07626	3.71	0.0009	0.18145	5.51106

Aucune des valeurs VIF ne sont trop grandes, et tous les coefficients sont significatifs.

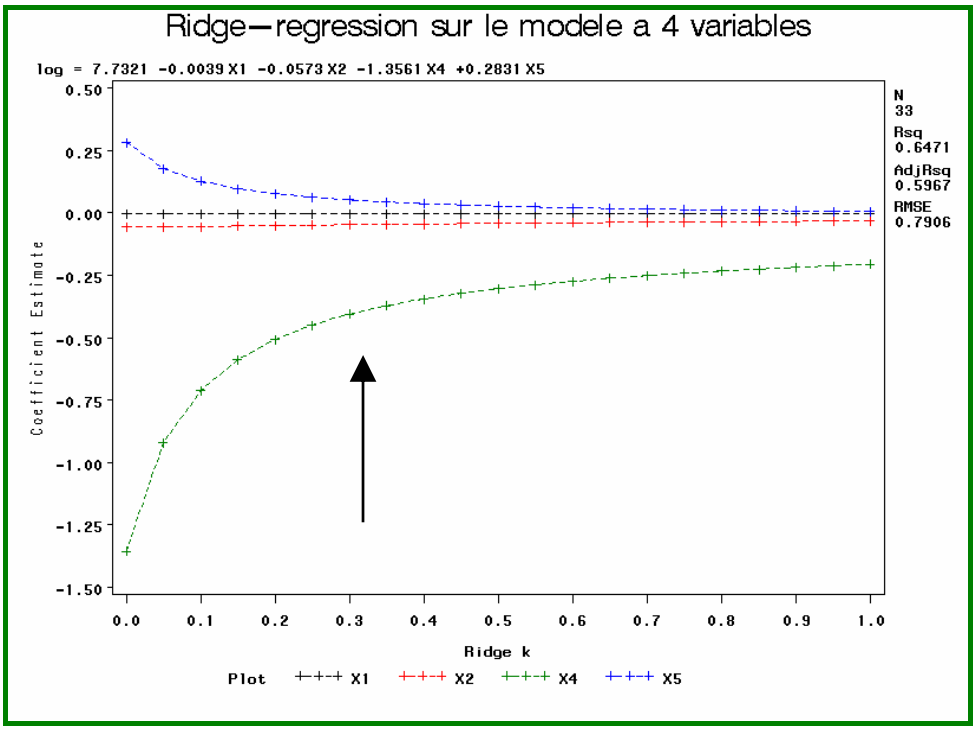
Collinearity Diagnostics (intercept adjusted)						
			Proportion de variation			
Nombre	Valeur propre	Index de condition	X1	X2	X4	X5
1	2.11413	1.00000	0.05758	0.01571	0.03506	0.03509
2	0.97576	1.47196	0.04243	0.87463	0.00581	0.00845
3	0.81585	1.60976	0.89080	0.10710	0.00847	0.01228
4	0.09426	4.73583	0.00920	0.00256	0.95067	0.94418

Dans la colonne « condition Index » (traduit dans la version française de SAS par « Index de condition ») il n'y a pas de grandes valeurs. Sur la 4ième et dernière ligne (c'est celle de « condition number »), en regardant les proportions de variance, on constate que les 2 variables X4 et X5 sont les « responsables » de la faiblesse de la 4ième valeur propre : on avait vu au chapitre 2 (§2.3.2) que c'est le couple de régresseurs le plus corrélé.

Regression RIDGE

Cette méthode, due à Hoerl et Kennard (1970), consiste à modifier (X'X) pour la rendre inversible. Pour cela on ajoute un terme constant k à la diagonale $(0 \leq k \leq 1)$. La solution des moindres carrés sera donc obtenue en inversant (X'X + kI) : les coefficients obtenus sont appelés « coefficients ridge ». On trace ensuite la variation des coefficients ridge en fonction de k : c'est la « Ridge Trace ». On détermine la valeur de k à partir de laquelle les coefficients se stabilisent : ce sera la valeur choisie.

```
title 'Ridge-regression sur le modele a 4 variables' ;
proc reg data=libreg.chenilles ridge = 0 to 1 by 0.05 outest =
coeff_ridge ;
LOG: model log=X1 X2 X4 X5 ;
plot / ridgeplot ;
run;
quit;
proc print data = coeff_ridge ;
run ;
```



Ici les coefficients ridge se stabilisent pour $k \cong 0.3$.
Les valeurs des coefficients sont alors lus dans la table coeff_ridge. Dans cette table, la première ligne est le modèle habituel, et la deuxième correspond à k=0, ce qui est le même modèle.

Pour _RIDGE_ = 0.3 ,
b1 = -0.003281930, b2 = -0.047532; b4 = -0.40556 ; b5 =0.05207.

Ridge-regression sur le modele a 4 variables

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	X1	X2	X4	X5	log
1	LOG	PARMS	log	.	.	0.79065	7.73214	-.003923681	-0.057343	-1.35614	0.28306	-1
2	LOG	RIDGE	log	0.00	.	0.79065	7.73214	-.003923681	-0.057343	-1.35614	0.28306	-1
3	LOG	RIDGE	log	0.05	.	0.81855	7.29248	-.003872789	-0.056112	-0.92225	0.17851	-1
4	LOG	RIDGE	log	0.10	.	0.85205	6.94148	-.003762661	-0.054392	-0.71231	0.12748	-1
5	LOG	RIDGE	log	0.15	.	0.87839	6.63795	-.003639039	-0.052591	-0.58856	0.09717	-1
6	LOG	RIDGE	log	0.20	.	0.89910	6.36685	-.003515143	-0.050826	-0.50692	0.07708	-1
7	LOG	RIDGE	log	0.25	.	0.91596	6.12061	-.003395569	-0.049135	-0.44894	0.06277	-1
8	LOG	RIDGE	log	0.30	.	0.93015	5.89464	-.003281930	-0.047532	-0.40556	0.05207	-1
9	LOG	RIDGE	log	0.35	.	0.94245	5.68578	-.003174656	-0.046017	-0.37181	0.04378	-1
10	LOG	RIDGE	log	0.40	.	0.95335	5.49170	-.003073668	-0.044589	-0.34473	0.03717	-1
11	LOG	RIDGE	log	0.45	.	0.96317	5.31059	-.002978676	-0.043242	-0.32248	0.03179	-1
12	LOG	RIDGE	log	0.50	.	0.97215	5.14097	-.002889300	-0.041972	-0.30381	0.02734	-1
13	LOG	RIDGE	log	0.55	.	0.98044	4.98162	-.002805140	-0.040772	-0.28788	0.02360	-1
14	LOG	RIDGE	log	0.60	.	0.98815	4.83153	-.002725802	-0.039638	-0.27411	0.02042	-1
15	LOG	RIDGE	log	0.65	.	0.99539	4.68982	-.002650914	-0.038566	-0.26204	0.01769	-1
16	LOG	RIDGE	log	0.70	.	1.00221	4.55575	-.002580129	-0.037550	-0.25137	0.01533	-1
17	LOG	RIDGE	log	0.75	.	1.00867	4.42864	-.002513130	-0.036586	-0.24183	0.01326	-1
18	LOG	RIDGE	log	0.80	.	1.01480	4.30792	-.002449626	-0.035670	-0.23324	0.01145	-1
19	LOG	RIDGE	log	0.85	.	1.02065	4.19309	-.002389353	-0.034800	-0.22545	0.00985	-1
20	LOG	RIDGE	log	0.90	.	1.02623	4.08369	-.002332073	-0.033971	-0.21835	0.00844	-1
21	LOG	RIDGE	log	0.95	.	1.03158	3.97931	-.002277566	-0.033181	-0.21183	0.00717	-1
22	LOG	RIDGE	log	1.00	.	1.03671	3.87960	-.002225634	-0.032428	-0.20581	0.00604	-1

4.5. Choix des régresseurs

Ce choix s'avère nécessaire en particulier si le nombre d'observations est petit par rapport au nombre de régresseurs, à cause du rang de X'X qui peut devenir plus petit que p. Ceci peut entrainer une instabilité des coefficients comme on l'a vu au paragraphe précédent.

Soit un modèle avec n observations et p régresseurs ; on sélectionne dans les cas suivants (cette liste est non exhaustive) :

1. n petit par rapport à p ;
2. colinéarité des régresseurs ;
3. choix d'un modèle plus simple pour la prévision (principe de PARCIMONIE).

(1 et 2 entraînent des problèmes d'inversion de X'X).

Proc REG permet ce choix par l'option « SELECTION = method », de l'instruction MODEL. Il n'y a pas de sélection dans SAS/INSIGHT.

4.5.1. Utilisation des sommes de carrés

La formule de base est : $SS_{Totale} = SS_{Modèle} + SS_{Erreurs}$

Rappel sur les somme de carrés apportés par un régresseur

Les sommes de carrés apportés par les régresseurs peuvent être obtenues par les options SS1 SS2 de l'instruction MODEL de REG, ce qui a déjà été vu dans le chapitre 2 (§2.4.1), ou bien par les « Type III Tests » dans SAS/INSIGHT.

- SS1(Xj) = somme des carrés apportée par la variable Xj introduite en séquence dans la régression, la régression contenant uniquement les variables qui la précèdent dans la liste de variables explicatives de l'instruction MODEL.
-
- SS2(Xj) = somme des carrés apportée par la variable Xj, lorsque l'ensemble des (p-1) autres régresseurs est déjà dans la régression. Ce sont les sommes de carrés données par la table « Type III Tests » de SAS/INSIGHT.

SS2(Xj) correspond au calcul du carré de la différence entre la valeur de Y estimée par la régression avec les p variables et celle estimée dans la régression à (p-1) variables, sans Xj. Pour le choix de régresseurs, le deuxième calcul d'apport de somme de carrés est le plus intéressant, car il ne dépend pas de l'ordre d'introduction des variables dans le modèle. On notera « $SS_{\text{apporté par j}}$ » cette quantité SS2(Xj).

Tests des apports à $SS_{Modèle}$ d'une variable

Les tests décrits dans le chapitre 2 (§2.4.3), ne sont pas faits par l'option SS2 de l'instruction MODEL de REG, mais sont donnés dans la table «Type III Tests » de SAS/INSIGHT.
Plus généralement, comme on l'a vu au §2.4.4, un modèle sans r variables est appelé modèle restreint par opposition au modèle complet à p variables.

- RRSS (Restricted Residual Sum of Squares) = Somme des carrés des résidus du modèle restreint
- URSS (Unrestricted Residual Sum of Squares)=somme des carrés des résidus du modèle complet.

La valeur de la statistique du test est $F = \frac{(RRSS - URSS)/r}{URSS/(n-p-1)}$.

Dans le cas d'une seule variable, r vaut 1, et donc en passant aux sommes de carrés du modèle :

$$F = \frac{(RRSS - URSS)/1}{URSS/(n-p-1)} = \frac{SS_{\text{Modèle complet}} - SS_{\text{Modèle sans j}}}{MSE}.$$

$$\boxed{F = \frac{SS_{\text{apporté par j}}}{MSE}}$$

Les tests de significativité de ces sommes de carrés sont donc réalisés à l'aide d'une statistique F, obtenue en divisant SS par la quantité MSE (Mean Square Error) du modèle avec constante contenant tous les régresseurs. De plus, la valeur de F associé à SS est aussi le carré du t de Student du coefficient de la variable j dans la régression à p régresseurs.

Exemple d'élimination progressive

On analyse les données « Processionnaire du pin » issu du livre de Tomassone et a*l.*.(1983). On va calculer les apports de sommes de carrés pour éliminer progressivement les variables à partir du modèle complet à 10 variables. On utilise les sorties de SAS/INSIGHT, qui permet dans le tableau « Type III Tests » de tester la validité de l'apport des sommes de carrés.

Modèle à 10 variables

Model Equation										
log	=	10.9984	-	0.0044 X1	-	0.0538 X2	+	0.0679 X3		
	-	1.2936 X4	+	0.2316 X5	-	0.3568 X6	-	0.2375 X7		
	+	0.1811 X8	-	1.2853 X9	-	0.4331 X10				

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	10	34.4662	3.4466	5.01	0.0008
Error	22	15.1299	0.6877		
C Total	32	49.5960			

Parameter Estimates							
Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept	1	10.9984	3.0603	3.59	0.0016	.	0
X1	1	-0.0044	0.0016	-2.85	0.0094	0.5327	1.8774
X2	1	-0.0538	0.0219	-2.46	0.0223	0.8400	1.1904
X3	1	0.0679	0.0995	0.68	0.5017	0.0239	41.8708
X4	1	-1.2936	0.5638	-2.29	0.0317	0.0624	16.0219
X5	1	0.2316	0.1044	2.22	0.0371	0.1066	9.3845
X6	1	-0.3568	1.5665	-0.23	0.8219	0.0170	58.7558
X7	1	-0.2375	1.0060	-0.24	0.8156	0.6064	1.6491
X8	1	0.1811	0.2367	0.76	0.4525	0.0693	14.4226
X9	1	-1.2853	0.8648	-1.49	0.1514	0.0895	11.1686
X10	1	-0.4331	0.7349	-0.59	0.5616	0.6057	1.6509

C'est un modèle globalement bon (F d'analyse de variance significatif), mais les coefficients des régresseurs X3 X6 X7 X8 X9 et X10 sont non significatifs . De plus, l'inflation de variance est plus grande que 10 pour les régresseurs X3 X4 X6 et X9. Le modèle n'est donc pas un « bon modèle » : il faut éliminer des régresseurs. Pour cela on effectue les tests sur les apports de sommes de carrés

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	5.5717	5.5717	8.10	0.0094
X2	1	4.1551	4.1551	6.04	0.0223
X3	1	0.3208	0.3208	0.47	0.5017
X4	1	3.6205	3.6205	5.26	0.0317
X5	1	3.3869	3.3869	4.92	0.0371
X6	1	0.0357	0.0357	0.05	0.8219
X7	1	0.0383	0.0383	0.06	0.8156
X8	1	0.4023	0.4023	0.59	0.4525
X9	1	1.5190	1.5190	2.21	0.1514
X10	1	0.2389	0.2389	0.35	0.5616

Les variables X3, X6, X7, X8, X9, et X10 ont un apport non significatif : on élimine la variable X6 dont l'apport est le plus petit.
Remarque : on peut vérifier ici que F(X6) = t^2(X6) → 0.0529 = $(-0.23)^2$ (cf. chapitre 1, §1.3.4)

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	5.5373	5.5373	8.40	0.0081
X2	1	4.1218	4.1218	6.25	0.0200
X3	1	0.9909	0.9909	1.50	0.2327
X4	1	4.0339	4.0339	6.12	0.0212
X5	1	3.5100	3.5100	5.32	0.0304
X7	1	0.0871	0.0871	0.13	0.7196
X8	1	0.3742	0.3742	0.57	0.4589
X9	1	1.8513	1.8513	2.81	0.1074
X10	1	0.2063	0.2063	0.31	0.5813

→ On élimine X7

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	6.3925	6.3925	10.06	0.0041
X2	1	4.0447	4.0447	6.36	0.0187
X3	1	0.9455	0.9455	1.49	0.2344
X4	1	4.9416	4.9416	7.78	0.0102
X5	1	4.7047	4.7047	7.40	0.0119
X8	1	0.3782	0.3782	0.60	0.4480
X9	1	1.7752	1.7752	2.79	0.1076
X10	1	0.2984	0.2984	0.47	0.4997

→ On élimine X8

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	7.0905	7.0905	11.34	0.0025
X2	1	4.1181	4.1181	6.59	0.0166
X3	1	1.1435	1.1435	1.83	0.1884
X4	1	5.0418	5.0418	8.06	0.0088
X5	1	4.5410	4.5410	7.26	0.0124
X9	1	1.4670	1.4670	2.35	0.1381
X10	1	0.4195	0.4195	0.67	0.4205

→ On élimine X10

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	6.6711	6.6711	10.81	0.0029
X2	1	4.2645	4.2645	6.91	0.0142
X3	1	0.9272	0.9272	1.50	0.2314
X4	1	4.7457	4.7457	7.69	0.0101
X5	1	4.3079	4.3079	6.98	0.0138
X9	1	1.4451	1.4451	2.34	0.1381

→ On élimine X3

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	5.8582	5.8582	9.32	0.0051
X2	1	4.0243	4.0243	6.40	0.0176
X4	1	6.2247	6.2247	9.90	0.0040
X5	1	5.6406	5.6406	8.97	0.0058
X9	1	0.5259	0.5259	0.84	0.3685

→ On élimine X9

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
X1	1	7.3067	7.3067	11.69	0.0019
X2	1	5.4683	5.4683	8.75	0.0062
X4	1	11.2389	11.2389	17.98	0.0002
X5	1	8.6123	8.6123	13.78	0.0009

On a obtenu un modèle dont les apports de toutes les variables sont tous significatifs : on arrête donc le processus.
La procédure réalisée donne donc comme modèle final Y = f(X1 X2 X4 X5); qui est le modèle déjà étudié et qui s'est révélé être un modèle « correct », vérifiant les suppositions de base sur les erreurs (§4.2.5), où 4 observations sont atypiques (§4.3.11), et dans lequel les 4 régresseurs n'ont pas de problème de colinéarité (§4.4.5).

4.5.2. Différentes méthodes basées sur les sommes de carrés

Méthode FORWARD (ascendante)

On introduit les variables une par une : on commence par un modèle à une variable, et on ajoute à chaque étape une variable. Les $SS_{Modèle}$ augmentent forcément (gain) et le principe est de faire entrer à chaque pas la variable qui apportera l'augmentation la plus significative de la somme des carrés du modèle. Donc, la variable qui est introduite est celle qui a $SS_{apporté\ par\ j}$ maximum, donc qui possède le F le plus grand, et significatif avec une probabilité par défaut associée au F de 0.5 : ce seuil s'appelle le seuil «pour entrer» SLE de REG.
Il y a au plus p modèles sélectionnés, qui sont affichés par ordre croissant de k (k=1 à L, L $\leq$ p.

Méthode BACKWARD (descendante)[28]

On part de la régression à p régresseurs, et on élimine à chaque pas la variable la moins significative, c'est-à-dire qu'on élimine la variable ayant $SS_{\text{apporté par j}}$ minimum, c'est-à-dire le F ou le t de Student le plus petit (probabilité par défaut associée au F de 0.10 : seuil «pour sortir» SLS de REG).
Il y a au plus p modèles sélectionnés, qui sont affichés par ordre décroissant de k (k=p à L, L $\geq$ 1.

Méthode STEPWISE (progressive)

C'st une combinaison FORWARD/BACKWARD : on effectue une sélection FORWARD, en laissant la possibilité de faire sortir du modèle à chaque pas une des variables devenue non significative (seuils de probabilité «pour entrer» 0.15, «pour sortir» 0.15 par défaut dans REG).

Remarque : les méthodes FORWARD, BACKWARD et STEPWISE ne donnent pas forcément le meilleur sous-ensemble à k variables. On peut le voir sur l'exemple ci-dessous, extrait de Brenot, Cazes et Lacourly (1975). On considère un modèle à 3 régresseurs. La figure 4.6 illustre graphiquement les liaisons dans R^n :

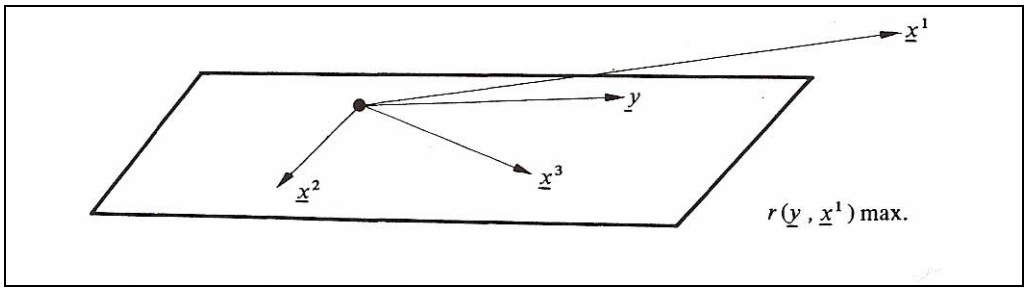


Figure 4.6 : représentation géométrique dans R^n de y et x1 x2 x3

Dans l'espace R^n, y, x1 et x3 sont coplanaires, et x1, qui n'appartient pas au plan, est telle que son coefficient de corrélation avec y est le plus fort des coefficients de corrélations de y avec les 3 régresseurs.
La meilleure régression avec 2 variables est donc y=f(x1). La meilleure régression à 2 variables est donc y=f(x2, x3).

Les méthodes de sélection donneront les choix successifs :

BACKWARD :	(x1, x2, x3)	→ (x2, x3)	→ x2 ou x3	
FORWARD :	x1	→ (x1, x2) ou (x1, x3)	→ (x1, x2, x3)	
STEPWISE :	x1	→ (x1,x 2) ou (x1, x3)	→ (x1, x2, x3)	→ (x2, x3)

Les modèles à 1 ou à 2 variables trouvés par ces méthodes sont différents, et ne sont pas forcément les meilleurs : BACKWARD trouve le meilleur modèle à 2 variables, mais pas celui à 1 ; FORWARD et STEPWISE trouvent le meilleur modèle à 1 variable mais pas celui à 2 ; la méthode STEPWISE effectue un pas de plus que FORWARD.

[28] C'est la méthode réalisée au §4.5.1.

Exemples de sélection STEPWISE

On sélectionne parmi les 10 variables candidates dans les données « Processionnaire du pin » issu du livre de Tomassone et *al.*(1983).

- Tout d'abord, on conserve les seuils par défaut SLE = SLS = 0.15 de la méthode.

```
title 'regression STEPWISE  ';
proc reg data =libreg.chenilles  ;
log10 : model log = x1--x10 / selection = stepwise;
run ;
quit;
```

Stepwise Selection: Step 1

Variable X9 Entered: R-Square = 0.3528 and C(p) = 17.6721

Analyse de variance					
Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	1	17.49867	17.49867	16.90	0.0003
Error	31	32.09735	1.03540		
Corrected Total	32	49.59603			

Variable	Résultat estimé des paramètres	Erreur std	Type II SS	Valeur F	Pr > F
Intercept	1.77374	0.65375	7.62205	7.36	0.0108
X9	-1.30538	0.31753	17.49867	16.90	0.0003

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable X1 Entered: R-Square = 0.4690 and C(p) = 11.2904

Analyse de variance					
Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	2	23.26290	11.63145	13.25	<.0001
Error	30	26.33313	0.87777		
Corrected Total	32	49.59603			

Variable	Résultat estimé des paramètres	Erreur std	Type II SS	Valeur F	Pr > F
Intercept	5.83839	1.69652	10.39561	11.84	0.0017
X1	-0.00353	0.00138	5.76422	6.57	0.0156
X9	-1.01284	0.31386	9.14074	10.41	0.0030

Bounds on condition number: 1.1525, 4.6099

Stepwise Selection: Step 3

Variable X2 Entered: R-Square = 0.5310 and C(p) = 8.8262

Analyse de variance					
Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	3	26.33301	8.77767	10.94	<.0001
Error	29	23.26301	0.80217		
Corrected Total	32	49.59603			

Variable	Résultat estimé des paramètres	Erreur std	Type II SS	Valeur F	Pr > F
Intercept	6.76034	1.68890	12.85278	16.02	0.0004
X1	-0.00352	0.00132	5.74447	7.16	0.0121
X2	-0.04486	0.02293	3.07012	3.83	0.0601
X9	-0.82500	0.31503	5.50133	6.86	0.0139

Bounds on condition number: 1.2705, 10.625

Stepwise Selection: Step 4

Variable X3 Entered: R-Square = 0.5799 and C(p) = 7.2958

Analyse de variance					
Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	4	28.76097	7.19024	9.66	<.0001
Error	28	20.83505	0.74411		
Corrected Total	32	49.59603			

Variable	Résultat estimé des paramètres	Erreur std	Type II SS	Valeur F	Pr > F
Intercept	9.54614	2.24151	13.49616	18.14	0.0002
X1	-0.00481	0.00145	8.13804	10.94	0.0026
X2	-0.04848	0.02217	3.55650	4.78	0.0373
X3	0.06911	0.03826	2.42796	3.26	0.0816
X9	-1.72321	0.58251	6.51184	8.75	0.0062

Bounds on condition number: 5.724, 52.201

Résultats :
La sélection se fait en 4 étapes. A chaque étape, des résultats globaux sur le modèle sont données (R^2, CP, analyse de variance). Pour vérifier si les régresseurs n'ont pas de colinéarité « pathologique », le «Condition Number » (cf. 4.4.3) est borné : ici par exemple au 4ième pas, Condition Number a une borne supérieure très grande (52.20) donc on peut soupçonner l'existence d'une colinéarité entre les 4 régresseurs.

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Étape	Variable entrée	Variable supprimée	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
1	X9		1	0.3528	0.3528	17.6721	16.90	0.0003
2	X1		2	0.1162	0.4690	11.2904	6.57	0.0156
3	X2		3	0.0619	0.5310	8.8262	3.83	0.0601
4	X3		4	0.0490	0.5799	7.2958	3.26	0.0816

Le modèle trouvé en 4 étapes est le modèle LOG=f(X9, X1, X2, X3).

- Puis on fixe les seuils SLE et SLS à 5% par les options SLE= 0.05 et SLS = 0.05.

```
title 'regression STEPWISE avec seuils à 0.05  ';
proc reg data =libreg.chenilles  ;
log10 : model log = x1--X10 / selection = stepwise SLE = 0.05
SLS = 0.05;
run ;
quit;
```

Dans ce cas, 2 pas seulement sont effectués et le modèle est LOG= f(X9, X1).

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection								
Étape	Variable entrée	Variable supprimée	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
1	X9		1	0.3528	0.3528	17.6721	16.90	0.0003
2	X1		2	0.1162	0.4690	11.2904	6.57	0.0156

Remarque
On peut raisonner soit avec les sommes de carrés comme on vient de le faire aux §4.5.1 et 4.5.2, soit avec $R^2 = \frac{SS_{Modèle}}{SS_{Total}}$, ce qui est équivalent compte-tenu de l'équation d'analyse de variance : $SS_{Totale} = SS_{Modèle} + SS_{Erreurs}$

D'où les autres méthodes présentées ci-dessous.

4.5.3. Amélioration de R²

Maximum R^2 Improvement (MAXR)

C'est une méthode qui procède par étape comme les précédentes. Elle tente de trouver le meilleur modèle au sens du R² pour chaque valeur k du nombre de régresseurs.
La méthode MAXR commence par choisir la variable donnant le plus grand R² (c'est à dire celle qui est la plus corrélée avec Y).
Puis est ajoutée celle qui provoque la plus grande augmentation du R².
Une fois ce modèle à 2 variables obtenu, tous les échanges possibles entre une des 2 variables présentes dans le modèle et une variable extérieure sont examinés, c'est-

à-dire que le R^2 de la régression est calculé, et l'échange qui est fait est celui qui fournit l'accroissement maximum de R^2.
La comparaison recommence alors avec 3 variables dans le modèle.
Ce processus continue jusqu'à ce qu'aucune permutation n'augmente R^2.

La différence entre les méthodes STEPWISE et MAXR est que toutes les permutations possibles sont évaluées dans MAXR **avant** le changement. Dans STEPWISE, seul le « moins bon » régresseur est retiré, sans vérifier si on pourrait ajouter la « meilleure » variable. En contre partie, MAXR demande évidemment beaucoup plus de calculs.

Minimum R^2 Improvement (MINR)

Il s'agit du même processus que le précédent sauf que la procédure d'´echange fait appel au couple de variables associé au plus petit accroissement du R^2. L'objectif est ainsi d'explorer plus de modèles que dans le cas MAXR et donc, éventuellement, de tomber sur un meilleur optimum.

4.5.4. Autres méthodes basées sur R^2 : RSQUARE et ADJRSQ

Ces méthodes ne fonctionnent pas par étapes. Elles affichent pour toute valeur de k, le *meilleur sous-ensemble* de k régresseurs au sens de R^2 ou R^2_{adj} (défini au chapitre 1 §1.2.5), ce que n'assurent pas les méthodes STEPWISE ou MAXR/MINR.
Elles demandent beaucoup plus de calculs car il faut examiner toutes les régressions possibles, mais les puissants moyens informatiques actuels rendent ces méthodes très rapides.

Pour k=1 à p, les modèles sont affichés dans l'ordre décroissant de R^2 (ou R^2_{adj}).
Pour limiter le volume des sorties quand le nombre de variables est grand, l'option BEST = q limite l'affichage aux q premiers modèles pour chaque valeur de k.

Ceci permet d'explorer rapidement les sous-ensembles de variables, mais les modèles sélectionnés, optimaux au sens R^2 (ou $R^2_{adj.}$), ne le sont pas forcément au niveau des données. Aussi il est recommandé d'afficher des critères supplémentaires de qualité comme les critères CP, AIC et BIC présentés ci-dessous.
Il faudra aussi valider les modèles, comme on l'a vu dans ce chapitre aux §4.2, §4.3, §4.4.

4.5.5. Coefficient CP de Mallows

Ce coefficient proposé par Mallows en 1973, est basé sur la recherche des régresseurs ayant le meilleur pouvoir prédictif, c'est à dire l'erreur totale moyenne la plus petite. Il permet ainsi de choisir entre plusieurs régressions différant à la fois par le nombre de régresseurs et la précision atteinte.
En effet, si on ajoute des régresseurs on diminue, en général, le biais des estimations mais on risque d'augmenter les variances des estimations et «l'erreur totale» moyenne, car on a pu ajouter des variables très liées aux variables initialement introduites.

La précision étant mesurée par MSE ou bien SSE, on calculera le coefficient ainsi :

- Si q < p , CP(q) =[SSE(q)/MSE] - [n - 2(q + 1)]
  - où SSE(q) = somme des carrés des erreurs de la régresssion avec q régresseurs et constante,
  - et MSE = précision s^2 calculée avec les p régresseurs (soit SSE(p)/(n-(p+1)).
- Si q = p, alors CP(p) = p+1.

Si le « bon » modèle est choisi, l'estimation est sans biais, et alors CP(q) est proche de (q+1) : voir plus de détails dans Daniel et Wood (1980).
Dans le cas où CP vaut (p+1), on a la même précision que la régression globale.

Sélection suivant le coefficient CP

En plus du R^2, REG affiche les sous-ensembles de régresseurs par ordre croissant de CP.
A partir du modèle complet, on choisit le premier sous-ensemble dont le CP approche la valeur (p+1): c'est le « meilleur » sous-ensemble de régresseurs.
On peut aussi faire le graphique de CP(q) en fonction de q (q=1 à p) et regarder la position de l'optimum ainsi que Mallows recommande:

Utilisation du coefficient CP dans une sélection de régresseurs

On demande à afficher le coefficient de chaque modèle obtenu par la méthode de sélection choisie, en ajoutant l'option CP à l'instruction MODEL de REG. Pour un nombre k donné de régresseurs, il est conseillé de choisir le sous-ensemble de k variables ayant une valeur de CP la plus grande.

4.5.6. Critères AIC et BIC

Ces critères ne sont pas des critères de sélection des régresseurs, mais des indicateurs de la qualité du modèle.

De manière générale, ces critères mesurent la qualité d'un modèle statistique bâti sur k paramètres sur un échantillon de taille n, à partir de la fonction de vraisemblance L.

Akaike Information Criterion (1969) : $AIC = -2\,Log(L) + 2k$

Sawa's Bayesian information criterion (1978) : $BIC = -2\,Log(L) + k\,Log(n)$

Le critère BIC est d'un autre critère utilisé nommé « critère de Schwartz ».

Dans le cas d'un modèle de régression à p régresseurs avec constante :

$$AIC = n\,Log\left(\frac{SSE}{n}\right) + 2(p+1)$$

$$BIC = n\,Log\left(\frac{SSE}{n}\right) + 2(p+3)q \quad \text{avec} \quad q = \frac{n\,s^2}{SSE}$$

Ces indicateurs sont utilisés comme la règle habituelle consistant à choisir le modèle ayant la meilleure précision (SSE petit).

4.5.7. Exemple de sélection RSQUARE

On sélectionne parmi les 10 variables candidates dans les données « Processionnaire du pin » issu du livre de Tomassone et *al*.(1983). On demande l'affichage des 2 meilleurs modèles pour chaque nombre de régresseurs par l'option BEST = 2.

```
title 'regression RSQUARE  ';
proc reg data =libreg.chenilles ;
log10 : model log = X1--X10 / selection = RSQUARE AIC BIC CP
BEST = 2 ;
run ;
quit;
```

Nombre dans le modèle	R-carré	C(p)	AIC	BIC	Variables du modèle
1	0.3528	17.6721	3.0848	3.7414	X9
1	0.2931	21.9804	5.9985	6.3390	X8
2	0.4690	11.2904	-1.4474	-0.3146	X1 X9
2	0.4249	14.4777	1.1911	1.8812	X1 X8
3	0.5368	8.4026	-3.9541	-2.0508	X1 X4 X5
3	0.5310	8.8262	-3.5382	-1.7348	X1 X2 X9
4	0.6471	2.4513	-10.9258	-6.1357	X1 X2 X4 X5
4	0.5799	7.2958	-5.1757	-2.2991	X1 X2 X3 X9
5	0.6577	3.6865	-9.9325	-4.1182	X1 X2 X4 X5 X9
5	0.6550	3.8819	-9.6724	-3.9701	X1 X2 X4 X5 X10
6	0.6764	4.3383	-9.7858	-2.3328	X1 X2 X3 X4 X5 X9
6	0.6764	4.3392	-9.7846	-2.3322	X1 X2 X4 X5 X6 X9
7	0.6864	5.6125	-8.8285	0.0994	X1 X2 X3 X4 X5 X8 X9
7	0.6848	5.7284	-8.6597	0.1627	X1 X2 X3 X4 X5 X9 X10
8	0.6925	7.1785	-7.4679	2.8386	X1 X2 X3 X4 X5 X8 X9 X10
8	0.6901	7.3519	-7.2110	2.9101	X1 X2 X3 X4 X5 X7 X8 X9
9	0.6942	9.0519	-5.6569	5.7796	X1 X2 X3 X4 X5 X7 X8 X9 X10
9	0.6942	9.0557	-5.6512	5.7806	X1 X2 X3 X4 X5 X6 X8 X9 X10
10	0.6949	11.0000	-3.7346	8.7654	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10

Dans le tableau des modèles, on constate que, parmi les modèles à 4 variables, celui qui a le meilleur R^2, et les plus petites valeurs de AIC et BIC est Y = f(X1 X2 X4 X5), qui est le modèle étudié aux paragraphes §4.2.5, §4.3.11 et §4.4.5. Par contre le critère CP est plus petit que celui 2ième modèle sélectionné Y = f(X1 X2 X3 X9), qui pourrait donc être également un modèle à étudier : c'est d'ailleurs le modèle final trouvé par la méthode STEPWISE (§4.5.2) avec les seuils par défaut de15 %.

Conclusion

Arrivé au terme de ce tutoriel nous voulons insister sur deux points non abordés par la technique de la régression : la qualité des données et les difficultés d'interprétation. Ces deux points sont du ressort du spécialiste du domaine d'études sur lequel le statisticien applique la régression.

La **qualité de l'information** apportée par les données (observations) intervient dans la validité et la robustesse d'un modèle de régression. Mais cette qualité n'est pas appréhendable par le statisticien-praticien.
Ce sont des connaissances *externes* à la statistique mais *internes* à l'étude qui doivent intervenir. Ces connaissances sont aussi indispensables pour déterminer le plan d'échantillonnage. Cette étape qui se situe au niveau de la collecte des données et donc en amont de l'analyse statistique des données mériterait à elle seule un long développement.

Une fois réglées toutes les difficultés reste un dernier point et non des moindres, **l'interprétation.** Prenons un exemple concret : parmi des enfants, on effectue une enquête permettant de mesurer l'étendue du vocabulaire et la taille de leurs pieds.
La corrélation entre ces deux variables est nettement significative ! Le bon sens permet d'éviter d'en tirer des conclusions aberrantes. Sous cette corrélation se cache l'influence de la variable âge.
Autre exemple, dans un état des U.S.A on a corrélé sur les 20 dernières années, le taux de criminalité et le taux de fréquentation dans les églises. Là aussi la corrélation obtenue est très élevée, mais le bon sens ne vient que peu en aide. La variable cachée est l'immigration italienne et irlandaise.

A la lumière de ces deux exemples, gardons-nous de toutes interprétations hâtives.

Avoir toujours à l'esprit que sous une corrélation peut se cacher l'effet d'une autre variable, ou d'un autre facteur.

On peut cependant utiliser le modèle identifié, s'il est correct, dans un but de *prévision* mais surtout pas dans un but de "*contrôle*" (action sur les variables explicatives dans l'espoir d'agir sur Y) ou d'*explication*. Sinon, on pourrait augmenter l'intelligence de nos enfants en augmentant la taille de leurs pieds!

L'erreur qui perdure dans la littérature, est de donner le nom de variable dépendante ou variable **expliquée** à Y et de variables indépendantes ou variables **explicatives** à X, ce qui amène à déduire logiquement qu'il existe une idée de cause à effet entre X et Y. « Mécaniquement », ce n'est pas l'objet de la régression.

La régression sur données d'observations ne permet pas de déduire une quelconque relation de cause à effet de X sur Y et/ou de Y sur X. Il faut d'autres pratiques méthodologiques pour expliquer la causalité qui peut avoir des formes multiples.

La liaison entre 2 variables X et Y peut se rencontrer dans 5 situations:

- X cause Y
- Y cause X
- X et Y inter-agissent l'une sur l'autre, problème de circularité, ou de rétro-action
- X et Y évoluent ensemble sous l'effet d'une même variable
- X et Y sont liées par hasard

La causalité peut être validée par “l’outil” régression que **si on peut faire des comparaisons** sur des groupes comparables. C'est un débat historique qui est de plus en plus d'actualité. Les économètres tentent de pallier la faiblesse des techniques de régression appliquées à des données d’observations, en essayant de se rapprocher des techniques expérimentales. Ils ont introduit la méthode des variables instrumentales, avec l’idée de pouvoir comparer des individus qui ne diffèrent que sur une seule dimension : le *traitement.* Ils ont également proposé de traiter des données issues « d’expériences naturelles » et « d’expériences contrôlées », c’est un premier pas vers des interprétations causales (voir BEHAGHEL (2006).

Terminons sur ce constat, les méthodes de régression sont des méthodes très puissantes, mais qui doivent être utilisées avec beaucoup de discernement et de prudence. En toute honnêteté il ne faut pas se contenter d'un seul modèle et d'une seule procédure REG, il faut en tester plusieurs.

C'est un travail d'explorateur et de détective. C’est ce que nous avons tenté de mettre en lumière.

ANNEXES

Annexe 1

Syntaxe simplifiée de la Procédure REG de SAS

La procédure REG est une procédure « interactive » permettant d'étudier plusieurs modèles en un seul appel de PROC REG. **On donne ici son utilisation pour l'étude d'un seul modèle.**

```
PROC REG options ;
        MODEL dépendante = régresseurs / options ;
        BY nom_var ;
        FREQ nom_var ;
        ID nom_var ;
        WEIGHT nom_var ;
        REWEIGHT expression / option ;
RUN ;
        TEST équation(s) ;
        RESTRICT équation(s) ;
        OUTPUT OUT = data_sas mot_clef = nom_var ;
        PLOT yvar*xvar='symbol' / options ;
QUIT ;
```

PROC REG options ;

DATA=NOMTAB data_set_option option commune à toutes les procédures

OUTEST = TAB permet de créer des tables SAS de résultats utiles, comme les coefficients estimés, et des résultats créés dans des options.

Autres options : ALL CORR NOPRINT SIMPLE USSCP

ALL Demande beaucoup d'impressions (induit l'option SIMPLE, USSCP, et CORR).
CORR Imprime la matrice de corrélation de toutes les variables du modèle.
NOPRINT Supprime les impressions.
SIMPLE Imprime somme, moyenne, variance, écart-type et somme des carrés non corrigée pour les variables utilisées dans REG.
USSCP Imprime les sommes de carrés non corrigées et la matrice des produits croisés pour toutes les variables utilisées dans REG.

PRESS Permet d'obtenir dans la table OUTEST le coefficient PRESS

RIDGE et PCOMIT pour les régressions Ridge et sur composantes principales (voir à la fin de cette annexe).

MODEL dépendante = régresseurs / options ;

dépendante : nom de la variable dépendante

régresseurs : liste des noms des p variables régresseurs

Remarque : on peut donner un label à l'instruction MODEL, label qui sera alors affiché dans les sorties.

Quelques options de l'instruction MODEL:

<u>*Sélection de régresseurs*</u>
Option sous la forme SELECTION = nom (où nom est un des mots-clefs de la liste ci-après) :

NONE	pas de sélection (choix par défaut)
FORWARD	sélection ascendante
BACKWARD	sélection descendante
STEPWISE	sélection progressive ascendante
MAXR, MINR	sélection basée sur gain maximum/minimum en R^2
RSQUARE, ADJRSQ	sélection du meilleur sous-ensemble au sens de R^2, R^2 ajusté
CP	sélection basée sur CP de Mallows

Autres options associées à SELECTION :

INCLUDE = n	inclure les n premières variables explicatives dans les modèles explorés (n<p)
SLE = valeur	seuil de significativité pour entrer
SLS = valeur	seuil de significativité pour rester
STOP = s	arrête l'exploration au meilleur sous-ensemble de s variables (avec s < p); STOP = p par défaut.
BEST = k	arrête l'exploration après k modèles.

Attention aux valeurs par défaut des seuils :
SLE = 0.50 en FORWARD et SLE = 0.15 en STEPWISE
SLS = 0.10 en BACKWARD et SLS = 0.15 en STEPWISE

Remarque : les différentes valeurs du critère de sélection choisi sont stockées dans la table OUTEST, où on trouve aussi les 2 variables :
_IN_ nombre de régresseurs hors constante
_P_ nombre de régresseurs y compris la constante si elle existe dans le modèle

<u>Autres options de l'instruction MODEL:</u>

Définir un modèle sans constante : NOINT

Afficher:
- des résultats complémentaires pour les observations :

- P (prévisions) CLI CLM (intervalles de prévision à 95 % individuels et sur la moyenne) R(résidus) INFLUENCE (indices de détection des observations influentes)
- des coefficients: DW (Durbin-Watson), CP (Cp de Mallows), BIC, AIC, etc.;
- les sommes de carrés SS1 SS2 (carrés de type I ou II)

Diagnostiquer des problèmes particuliers:
- heteroscédasticité: SPEC (et ACOV)
- colinéarité: TOL VIF et COLLIN COLLINOINT

Instructions BY FREQ ID WEIGHT :

Ce sont des instructions communes à toutes les procédures. En particulier, WEIGHT permet de définir une régression pondérée.

REWEIGHT expression / WEIGHT = valeur ;

Cette instruction permet de redéfinir les poids, et en particulier d'omettre une observation de la régression.

expression est une comparaison sur une variable (on peut utiliser la variable OBS**.** qui contient le numéro de l'observation),
WEIGHT = valeur donne cette valeur de poids aux observations vérifiant l'expression .

Exemple : pour supprimer l'observation numéro 20, on écrit
REWEIGHT obs. = 20 / WEIGHT = 0 ;

TEST equation(s) ;

Cette instruction permet de tester une ou des hypothèses sur les estimations des paramètres (les équations doivent être séparées par des virgules). Chaque équation est une fonction linéaire formée de coefficients et de noms de variables (ici, INTERCEPT est le nom de la constante).

Exemples: TEST X1 = 0 , INTERCEPT = 0 ; *tester ($\beta 1=0$) et ($\beta 0=0$)*
TEST X3-X4 = 0 ; *tester ($\beta 3= \beta 4$)*

RESTRICT equation(s);

Elle permet de fixer des contraintes sur les coefficients, avec des équations identiques à TEST.

Exemple: RESTRICT X1-X3 = 0 ; *modèle avec contrainte ($\beta 1 = \beta 3$)*

OUTPUT OUT = nomtab mot_clef = nom_var ;

Cette instruction permet de créer une table SAS contenant certaines des variables créées par la régression. Ce tableau contiendra aussi les variables du modèle (réponse et régresseurs).

Liste des mots_clefs pour les variables créées par la régression
(ces mots_clefs sont utilisables aussi pour les instructions PLOT et PRINT, sauf PRESS):

PREDICTED (ou P) valeur prédite
L95M U95M limites des intervalles à 95% sur la moyenne des valeurs prédites
L95 U95 limites des intervalles à 95% sur une valeur prédite
STDP écart-type de la valeur moyenne prédite
STDI écart-type de la valeur prédite

RESIDUAL (ou R) résidu
STDR écart-type du résidu
NQQ quantile normal (pour le dessin QQPLOT)
STUDENT résidu studentisé interne
RSTUDENT résidu studentisé externe

H levier
PRESS coefficient Press (individuel)
COOKD DFFITS COVRATIO mesures d'influence des observations

PLOT Yvar1*Xvar1='s' Yvar2*Xvar2='s' / options ;

Cette instruction permet de tracer des graphiques en désignant les variables ordonnée Yvar et abscisse Xvar, et le symbole associé.

Différentes options sont possibles pour définir les caractéristiques des graphiques (cf. la documentation SAS).

Attention : on peut utiliser une des variables créées par la régression, définies plus haut dans la liste des mots-clefs de l'instruction OUTPUT, **à condition de faire suivre son nom par un point** : par exemple **P.** ou **R.**
On peut aussi utiliser la variable **OBS.** pour désigner le numéro de l'observation.

Remarque : l'option RIDGEPLOT permet de tracer le dessin des coefficients RIDGE (voir plus loin la description de l'option RIDGE).

PRINT mots-clefs;

Cette instruction permet d'imprimer certaines des variables créées avec la liste des mots-clefs vue plus haut.

Options RIDGE et PCOMIT des instructions PROC REG ou MODEL

On peut effectuer une régression Ridge ou une régression sur composantes principales, par une option de PROC REG ou de MODEL. La procédure travaille alors sur les données centrées (l'option NOINT est ignorée).

RIDGE = liste liste est une liste de valeurs qui peut être définie par la syntaxe *kd to kf by p,* où l'intervalle de variation du coefficient ridge est [kd,kf], la variation se faisant par pas de p.

Chaque valeur donne une estimation des coefficients Ridge, qui est placée dans une table SAS à définir avec l'option *OUTEST = table* de PROC REG. La colonne _TYPE_ indique quelle méthode on a employé : pour la méthode Ridge, _TYPE_=RIDGE, et les valeurs de la liste sont stockées sous le nom de variable _RIDGE_. On trouve ensuite les valeurs des coefficients Ridge de chaque régresseur.

PCOMIT = k k est un entier positif ou nul.

La procédure calcule alors les paramètres estimés en utilisant les composantes principales à l'exclusion des k dernières ; L'estimation des coefficients est placée dans une table SAS à définir avec l'option *OUTEST = table* de PROC REG, avec ici _TYPE_ = IPC.

Remarque : k peut aussi être une liste d'entiers non négatifs, pour permettre de faire plusieurs essais d'élimination de composantes.

Autres options de PROC REG tilisables en association avec les options RIDGE et PCOMIT :

OUTSTB pour avoir les estimations standardisés des coefficients estimés par RIDGE ou IPC ;

OUTSEB pour avoir les erreurs standardisées des coefficients ;

OUTVIF pour avoir les Variance Inflation Factor des coefficients.

Dessin « Ridge Trace »

Le dessin des coefficients Ridge en fonction des valeurs du paramètre (définies par l'option RIDGE = liste) est obtenu par l'instruction PLOT avec l'option RIDGEPLOT, à condition que les coefficients soient stockés dans une table par l'option OUTEST. On écrit alors simplement l'instruction :

PLOT / RIDGEPLOT ;

Annexe 2

Mode d'emploi très succinct de SAS/INSIGHT

Le module de SAS/Insight est à la fois un tableur un grapheur et un analyseur. Il permet de faire de l'Analyse Exploratoire des Données et de l'analyse confirmatoire dans l'esprit de TUKEY. Il est particulièrement bien adapté à la régression linéaire couplée à l'AED, grâce à ses possibilités de visualisation et d'interactivité.

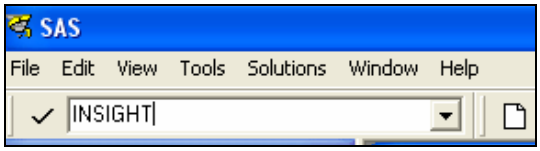
Tableau des Grandes Fonctions de SAS/INSIGHT

Analyse Exploratoire Techniques	Méthodes	Analyse confirmatoire
Visualisation graphiques 1D, 2D, 3D	**Analyse des distributions** - paramétrique - non paramétrique - QQ plots	**Ajustement par modèles** Régression linéaire, Analyse de - variance, - covariance GLM (MLG) Rég. Logistique Logit Probit
Exploration des données Rechercher, trier et éditer Identifier / données Brosser (Brushing et Slicing) Colorier , Marquer / critère Supprimer, cacher, exclure des calculs Animation etc.	**Analyse Composantes Principales** Rotation axes Analyse Canonique des Corrélations Analyse Discriminante	
Ré-expression des données	Plus de 20 fonctions mathématiques ou statistiques	

Nous ne présentons que quelques manipulations essentielles de SAS/INSIGHT. Pour une présentation plus complète voir l'ouvrage de DESTANDAU S. & LE GUEN M.

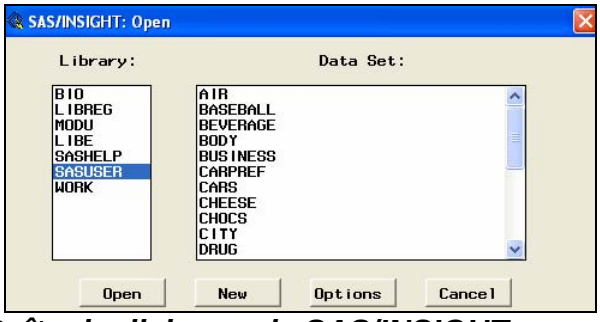
Le lancement de SAS/INSIGHT

➔ dans la barre de commande de SAS**,** taper **: INSIGHT** puis **entrée**



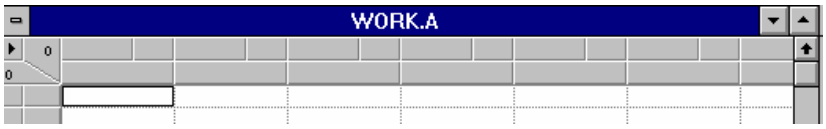
1. *Si la table de données n'existe pas encore*

Dans la boîte de dialogue, cliquer sur le bouton-poussoir **New**



Boîte de dialogue de SAS/INSIGHT

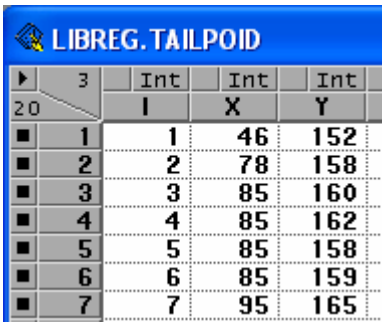
Un tableau de données vide s'ouvre. Saisissez vos données.

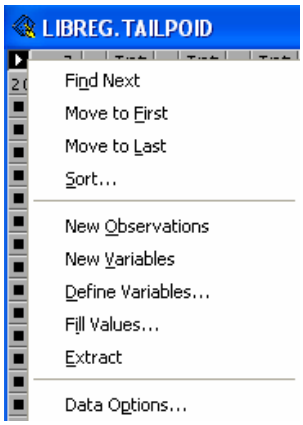


2. *Si la table existe*

Dans la boîte de dialogue sélectionner la bibliothèque : « **Library** » et la table SAS « **Data set** », et cliquer sur le bouton **Open**.

Affichage de la table SAS dans un tableur (cf. ci-dessous écran à gauche) et menu déroulant (cf. écran à droite).

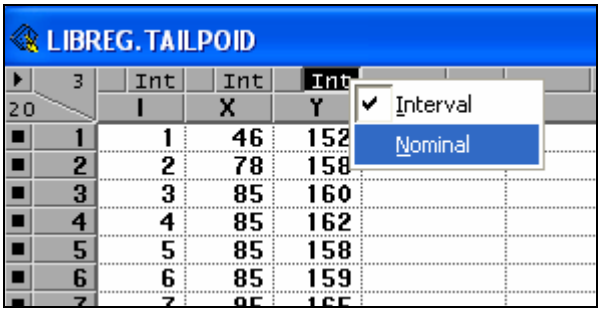




La Table TAILPOID a 3 variables et 20 observations indiqué par  .
En cliquant sur la petit flèche en haut à gauche le menu déroulant –pop menu ou encore menu contextuel- s'affiche avec les actions possibles sur le tableur.

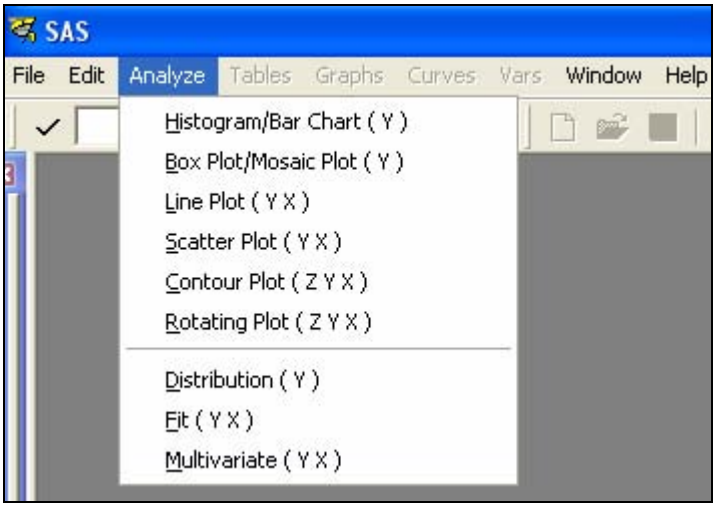
Rôle statistique des variables dans SAS/INSIGHT

Dans SAS/INSIGHT toute variable SAS définit en **caractère** est forcément une variable **nominale.** Par défaut une variable **numérique** n'est pas nominale, elle est **d'intervalle**. C'est à l'utilisateur de choisir le type d'échelles de mesures (Interval/Nominal) souhaité, en cliquant et cochant la zone au dessus du nom de la variable



Ce rôle statistique déterminera les types de graphiques à 1 dimension, 2 dimensions ou 3 dimensions et les types d'analyses.

Menu principal de SAS/INSIGHT



Graphiques standard en SAS/INSIGHT

- Graphiques pour les variable nominales : Bar Chart (1D) , Mosaic Plot (1D)
- Graphiques pour les variables d'intervalle : Histogram (1D) , Box Plot (1D) , Line Plot(2D), Scatter Plot (2D) , Contour Plot (3D), Rotating Plot (3D).

Pour réaliser un graphique il y a 2 possibilités : en utilisant les options par défaut, ou en passant par une boîte de dialogue pour modifier les options par défaut.
C'est un principe général dans SAS/INSIGHT.

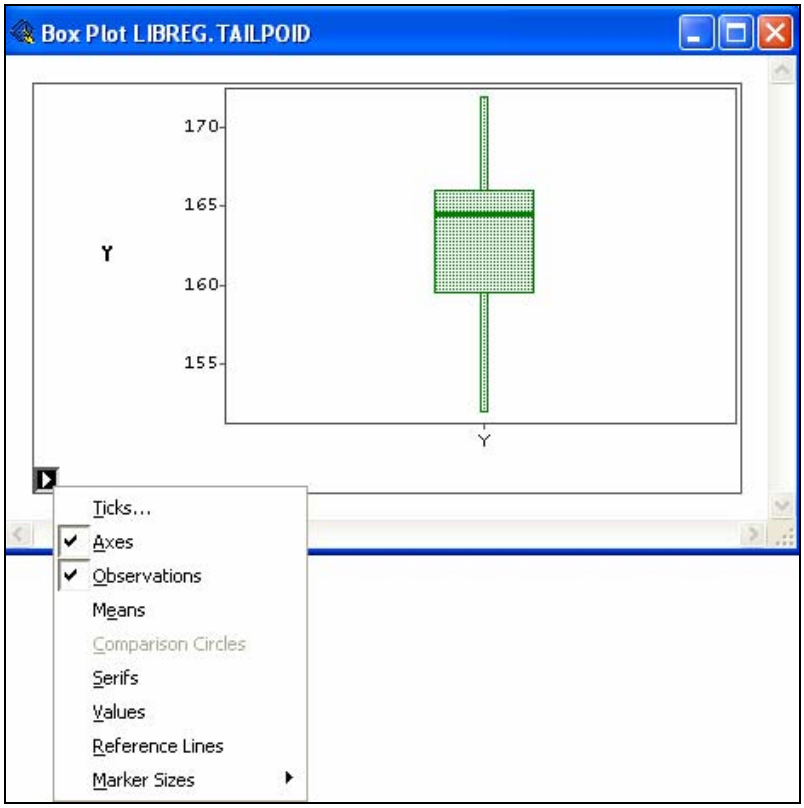
<u>Choix 1 – avec options par défaut</u>

Dans le tableur :

➔ **Cliquer** sur le nom de la variable d'intervalle Y

➔ menu : **Analyze# Box Plot/Mosaic Plot(Y)**

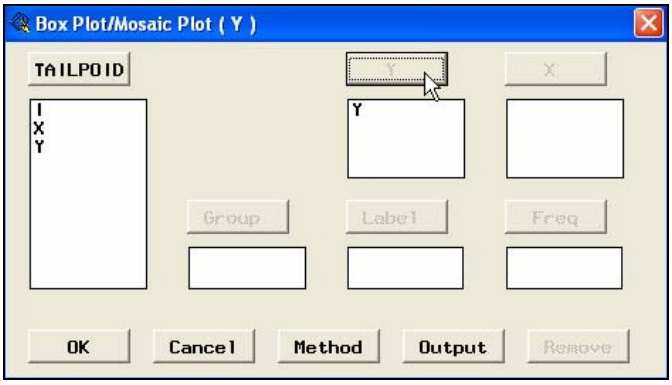
L'affichage est immédiat avec les options par défaut.



Sur le graphique, en cliquant sur la flèche en bas à gauche un menu déroulant s'affiche pour modifier les options. Par exemple, ajouter la moyenne avec **Means**, ajouter les valeurs des quantiles avec **Values** etc.

<u>Choix 2 – avec options modifiables</u>

➔ menu : **Analyze# Box Plot/Mosaic Plot(Y)**



Dans la boîte de dialogue qui s'affiche sélectionner la variable Y (dans la liste à gauche) puis cliquer sur le bouton-poussoir Y, pour que la variable choisie Y soit sélectionnée. Les boutons poussoirs **Method** et **Output** permettent de modifier les options par défaut.

Pour plus d'information sur les graphiques voir les articles en ligne sur les sites internet donnés en fin de cette annexe (page 354)

Les Analyses Statistiques avec SAS/INSIGHT

➔ menu : **Analyze # Distribution**
Etude d'une distribution (équivalent à Proc Univariate)

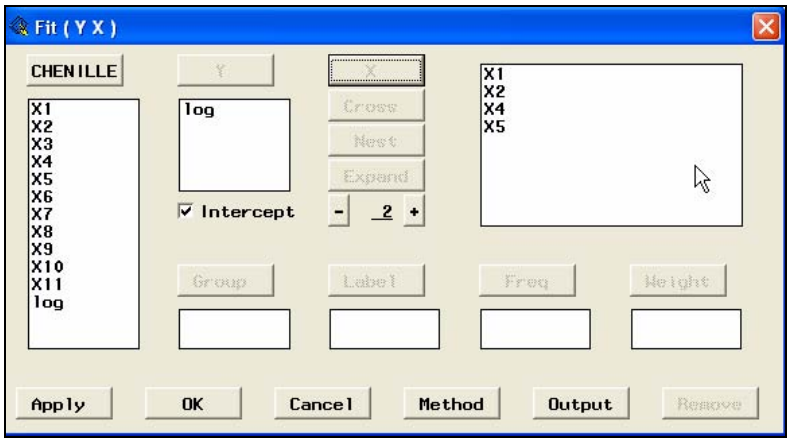
➔ menu : **Analyze # Fit** ➔ Analyse de régression linéaire, GLM , Régression logistique, Probit, Logit, ANOVA

➔ menu : **Analyze # Multivariate** ➔ Analyse en Composantes Principales, Analyse canonique, Analyse discriminante.

Même principe que pour les graphiques :

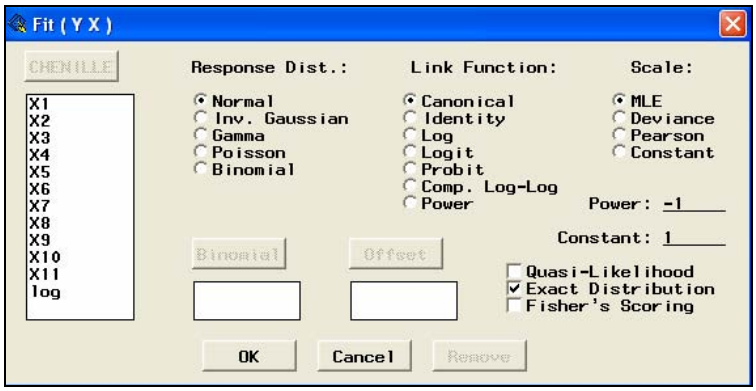
- Soit on sélectionne la ou les variables puis le menu **Analyze #.**.... Les analyses sont réalisées avec les options par défaut,
- Soit on choisit le menu **Analyze** #.... , une boîte de dialogue s'affiche pour sélectionner la ou les variables, et choisir les nouvelles options.

Exemple de Régression linéaire sur la Table SAS : Chenille (processionnaire du pin du §2.3.1.)



➔ Sélectionner dans la liste de gauche la variable Réponse (Log) puis cliquer sur le bouton de rôle Y, idem pour les variables régresseurs (X4,X2,X4,X5), en cliquant sur le bouton de rôle X. Si on veut la constante 1 (β_0) dans le modèle, cocher **Intercept**.

➔Cliquer sur le bouton poussoir Method :



Pour la régression linéaire les options à cocher sont :

- **Response Dist** : Normal
- **Link Function** : Canonical
- **Scale** : MLE

Impression et Sauvegarde

Nous présentons seulement quelques possibilités, pour imprimer un ou des éléments affichés, puis les sauvegarder dans un fichier externe, et enfin les insérer dans un document Word.

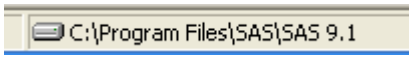
Pour imprimer

➔Sélectionner avec la souris, le graphique ou le tableau à imprimer, ou choisir
Menu : **Edit# Windows # Select all** pour sélectionner tous les éléments affichés
➔ **File # Print**

Pour sauvegarder les résultats graphiques ou tableaux dans un fichier

➔Sélectionner avec la souris la bordure du graphique ou le tableau à sauvegarder, ou choisir
Menu : **Edit# Windows # Select all** pour sélectionner tous les éléments affichés
➔ **File # save # Graphics File…**
puis renseigner la boîte de dialogue en choisissant par exemple le format .bmp et en suffixant le nom du fichier par .bmp (SAS ne le fait pas).

Le fichier sera sauvegardé dans le répertoire courant qui est affiché en bas de l'écran de SAS



Pour modifier l'emplacement, il suffit de double cliquer dessus et de changer le répertoire (fenêtre *Change Folder*).

Pour insérer un fichier externe .bmp dans Word

Dans un document Word : ➔ **insertion # image # à partir d'un fichier**…..

Pour plus d'information sur les graphiques

Consulter les articles en ligne sur les sites internet :

DESTANDAU S., LADIRAY D., M. LE GUEN, (1999), « AED mode d'emploi », Courrier des Statistiques, INSEE, n° 90, http://www.insee.fr/fr/ffc/docs_ffc/cs90e.pdf

LE GUEN M. (2001), La boîte à moustaches de Tukey, un outil pour initier à la Statistique, Statistiquement Vôtre, n° 4, 14 pages.
http://matisse.univ-paris1.fr/leguen/leguen2001b.pdf

LE GUEN M. (2004), « L'Analyse Exploratoire des Données et SAS/Insight, Visualisation Dynamiques des Données », Cahiers de la Maison des Sciences Economiques, Matisse, Série rouge, n°2004.01,13 pages,
ftp://mse.univ-paris1.fr/pub/mse/cahiers2004/R04001.pdf

CONFAIS J. & LE GUEN M., (2003), Graphiques conventionnels et Graphiques moins conventionnels. Importance de la visualisation Interactive, Document de travail ISUP-MATISSE, n°2003, 21 pages.
http://matisse.univ-paris1.fr/doc2/leguen1490.pdf

Annexe 3

Statistiques relatives à l'analyse de la variance

Statistique	Formule	Signification
Mean Square	$MS = \frac{SS}{DF}$	C'est le rapport d'une somme des carrés des écarts (SS) divisée par le nombre de degrés de liberté (DF). • pour SS model DF=p • pour SS error DF=n-p-1 La statistique Mean Square Error donne l'estimation s^2 de la vraie valeur inconnue de la variance des erreurs σ^2.
F Value	$F = \frac{\text{MS Model}}{\text{MS Error}}$	Statistique de Fisher-Snedecor pour tester si tous les paramètres β sont nuls.
Prob>F	$\text{Probf}(\text{F Value}, \text{ndf}, \text{ddf})$	C'est la *p-value* ou niveau de significativité du test, associée à F Value. La *p-value* est calculée en utilisant la fonction SAS : Probf. Probf : fonction SAS de la fonction de répartition d'une variable de Fisher-Snedecor ndf : nombre de degrés de liberté du numérateur de F Value ddf : nombre de degrés de liberté du dénominateur de F Value.
Root MSE	$\sqrt{\text{MS Error}}$	Standard Deviation, soit l'écart moyen résiduel. C'est l'estimation de "s", l'écart-type des erreurs
R-square	$R^2 = \frac{\text{SS Model}}{\text{SS Total}} = \frac{\text{SS Total} - \text{SS Error}}{\text{SS total}}$	F et Rsquare (R^2) sont liés par la relation $F = \frac{n-p-1}{p} * \frac{\text{Rsquare}}{1-\text{Rsquare}}$
Dep Mean	$\overline{Y} = \sum_{i=1,n} Y_i / n$	moyenne de la variable réponse Y

Statistique	Formule	Signification
Adj R-sq	$1-\frac{((n-\text{intercept})(1-R^2))}{n-p}$	R^2 ajusté en fonction du nombre de régresseurs du modèle. Intercept=0 s'il n'y a pas de constante[29] sinon intercept =1.
CV	$=\frac{\text{Root MSE}}{\text{Dep Mean}}*100$	Coefficient de variation exprimé en %

Statistiques sur les paramètres

Pour chaque paramètre β_j, SAS donne : l'estimation du paramètre avec son erreur-type, le test de l'hypothèse nulle ($\beta_i = 0$) et la *p-value* associée.

Remarque : La variable notée Intercept correspond à la variable constante $X_0 = 1$.

Statistique	Formule	Signification
Estimate	solution de : $(X'X)*b=(X'Y)$	estimation du paramètre β_j
Standard Error	$\sqrt{(X'X)_{ii}^{-1}.MSE}$	erreur-type de l'estimateur du paramètre β_j calculé à partir du $j^{\text{ième}}$ élément de la diagonale de la matrice $(X'X)^{-1}$
T for H_0 : Parameter=0	$T=\frac{\text{Estimate}}{\text{Std Error of Estimate}}$	statistique T de Student pour tester l'hypothèse nulle: H_0 : paramètre β_j =0 contre H_a : paramètre $\beta_j \neq 0$ Remarque $\boxed{T^2 = \text{Fvalue partiel}}$
Prob > !T!	Probt(T,df)	C'est la *p-value* ou niveau de significativité du test de Student. La *p-value* est calculée en utilisant la fonction SAS : Probt. Probt : fonction SAS de la fonction de répartition d'une variable de Student à df (Degree of Freedom) degrés de liberté.

[29] S'il n'y a pas de constante b_0 à l'origine, les statistiques relatives à l'analyse de la variance n'ont pas la même interprétation.

Annexe 4

Relations entre la loi normale et les statistiques de lois

Chi2, T de Student et F de Fisher-Snedecor

Normale si $\boxed{X \sim N(0,1)}$

Chi2 alors $\boxed{Z = X^2}$ suit une loi de χ^2 à 1 degré de liberté (ddl)

et $\boxed{Z = \sum_{i=1,n} X_i^2}$ suit un χ^2 à n ddl, si les X_i sont **indépendants** et N(0,1)

T de Student à n ddl

Si Z suit une loi de χ^2 à n ddl et si Z est indépendant de X

alors $\boxed{T = \frac{X}{\sqrt{\frac{Z}{n}}}}$ suit une loi de Student à n ddl.

F de Fisher-Snedecor

Si Z_1 et Z_2 sont des variables aléatoires **indépendantes** suivant chacune une loi de
χ^2 à $\nu 1$ et $\nu 2$ ddl

alors $\boxed{F = \frac{\frac{Z_1}{\nu_1}}{\frac{Z_2}{\nu_2}}}$ suit une loi de Fisher-Snedecor à (ν_1, ν_2) ddl.

Annexe 5

Construction d'un QQ-Plot

Ce graphique permet une visualisation de l'adéquation à une loi. Dans le cas de la loi normale, il est appelé « droite de Henry »

Principe de la droite de Henry

1. Soit une variable X dont on veut vérifier l'adéquation à une loi normale (m,σ)

On dispose de n observations de X : (x_i) pour i =1 à n .

On note F(x_i) la fonction de répartition empirique en (x_i) :

$$F(x_i) = \text{prob}(X \le x_i) = \text{prob}\left(\frac{X-m}{\sigma} \le \frac{x_i - m}{\sigma}\right)$$

Soit ϕ la fonction de répartition de la loi normale (0,1) : on peut trouver une valeur u_i de même fonction de répartition : $\exists u_i$ tel que $\phi(u_i) = F(x_i) \Leftrightarrow u_i = \phi^{-1}(F(x_i))$.

2. Si X suit une loi normale (m,σ) alors :

$$F(x_i) = \text{prob}(X \le x_i) = \text{prob}\left(\frac{X-m}{\sigma} \le \frac{x_i - m}{\sigma}\right) = \phi\left(\frac{x_i - m}{\sigma}\right)$$

$$\Rightarrow u_i = \phi^{-1}(F(x_i)) = \phi^{-1}\left(\phi\left(\frac{x_i - m}{\sigma}\right)\right) = \frac{x_i - m}{\sigma}$$

Donc les points (x_i, u_i) sont alignés sur la droite d'équation u_i = (x_i - m) / σ .

3. En pratique, on ordonne les valeurs x_i : on note $x_{(i)}$ les valeurs ordonnées.

Bien sûr, on a alors F($x_{(i)}$) = i /n .

Le « nuage de points » à représenter est donc défini ainsi : [$u_i = \phi^{-1}$ (i/n) ; $x_{(i)}$]

Attention : pour faire les mêmes représentations que SAS, on mettra les valeurs de X *en ordonnées*.

La droite d'équation **Y = (X-m)/σ** est celle dont les points du nuage doivent se rapprocher en cas d'adéquation à la loi normale : on la représente donc également sur le même plan.

Remarque : Proc UNIVARIATE du module SAS/Base, ainsi que le menu « Distribution » de SAS-INSIGHT, utilise un calcul légèrement différent de celui exposé ici :

On cherche $u_i = \phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right)$ où r_i est le rang de l'observation i ($r_i = i$ en général).

Ceci permet en particulier de ne pas « perdre » les points extrêmes.

Généralisation

Si on veut visualiser l'adéquation à une autre loi que la loi normale, il suffit de connaître la fonction de répartition G de cette loi, et que celle-ci soit inversible.

On remplace alors ϕ par G dans les formules du §2.

QQ-Plot avec SAS

SAS permet de représenter des QQ-Plot pour les lois suivantes :

Normale, LogNormale, Exponentielle et Weibull.

Dans le module SAS/Base, la procédure UNIVARIATE possède une instruction QQPLOT (voir la documentation SAS pour son utilisation un peu complexe).

Dans SAS/INSIGHT, on les trouve dans le menu « Distribution », rubrique « Graphs » → « QQ-Plot ». Pour tracer la droite de référence, dans le menu « Curves » demander « QQ ref line ».
Lorsque l'on a exécuté une régression linéaire avec le menu « Fit », on peut ajouter aux sorties standards un graphique QQ-Plot appelé « Residual Normal QQ » dans le menu « Graphs » (penser à cocher « Reference lines » dans le menu contextuel du graphique pour tracer la droite).

Bibliographie

AKAIKE, H., (1969), « Fitting Autoregressive Models for Prediction », *Annals of the Institute of Statistical Mathematics*, 21, 243 - 247.

ARMATTE M., (1995), *Histoire du modèle linéaire, Formes et Usages en Statistique et Econométrie jusqu'en 1945*, Thèse de Doctorat, EPHE le 24/01/1995

ANSCOMBE F., « Graphs in Statistical Analysis », *The American Statistician*, February 1973, vol.27, n°1, p17-21.

BEHAGHEL L. ,(2006), *Lire l'économétrie*, Collection Repères, Editions La Découverte.

BELSLEY D.A., KUH E., WELSH R.E,. (1980), *Regression diagnostics*, Wiley.

BENZECRI J.P., BENZECRI F., (1989), « Calculs de corrélation entre variables et juxtaposition de tableaux », *Les cahiers de l'analyse de données*, 1989, n° 3, pp347-354.

BRENOT J., CAZES P., LACOURLY N. ,(1975) « Pratique de la régression : Qualité et protection », *Cahiers du BURO n° 23*, pp 1-81.

CAZES P., (1975), « Protection de la régression par utilisation de contraintes linéaires et non linéaires », *Revue de statistique appliquée*, volume XXIII, numéro 3.

CAZES P., (1976), « Régression par boule et par l'Analyse des Correspondances », *Revue de statistique appliquée*, Vol XXIV n°4, pp5-22.

CHATTERJEE S., HADI A. S., (1988), *Sensitivity analysis in linear regression*, Wiley.

CHOW, G.C., (1960), « Tests of Equality between Sets of Coefficients in Two Linear Regressions », *Econometrica*, 28, 591-605.

CLEVELAND W. S., (1979), « Robust Locally Weighted Regression and Smoothing Scatterplots », *Journal of the American Statistical Association*, Vol. 74, pp. 829-836.

CLEVELAND W. S., (1993), *Visualizing Data*, Hobart Press, Summit, New Jersey, USA 1993.

CLEVELAND W. S., (1994), *The Elements of Graphing Data*, Hobart Press, Summit, New Jersey, USA 1994.

CONFAIS J., LE GUEN M., (2003), « La régression linéaire sous SAS », *Document de travail n°F9605 de la Direction des Statistiques Démographiques et Sociales de l'INSEE*

CONFAIS J., LE GUEN M., (2003), « Graphiques conventionnels et Graphiques moins conventionnels. Importance de la visualisation Interactive », *Document de travail ISUP-MATISSE*, n°2003, 21 pages.
http://matisse.univ-paris1.fr/doc2/leguen1490.pdf

COOK R.D., WEISBERG S., (1994), *An Introduction to Regression Graphics*, Wiley Series in Probability and Statistics.

DESJARDINS D., (1998), « Outliers, Inliers, and Just Plain Liars—New Graphical EDA + (EDA Plus) Techniques for Understanding Data », SUGI 26, SAS.
http://www2.sas.com/proceedings/sugi26/p169-26.pdf

DESTANDAU S., LE GUEN M., (1995), *Analyse exploratoire des données avec SAS/INSIGHT*, INSEE GUIDES N°7-8.

DANIEL, C.,WOOD, F. ,(1980), *Fitting Equations to Data*, Revised Edition, New York: John Wiley & Sons, Inc.

DESROSIÈRES A., (1993), *La politique des grands nombres, histoire de la raison statistique*, Editions la Découverte

DESTANDAU S., LADIRAY D., M. LE GUEN, (1999), « AED mode d'emploi », *Courrier des Statistiques, INSEE*, n° 90, http://www.insee.fr/fr/ffc/docs_ffc/cs90e.pdf

DRAPER N.R., SMITH H., (1966), *Applied regression analysis*, Wiley.

ERICKSON B.H. & NOSANCHUK T.A., (1995-2d édition), *Understanding Data*, Open Université Press, 381 pages.

ERKEL-ROUSSE H., (1990), «Détection et effets de la multicolinéarité dans les modèles linéaires ordinaires », *Document de travail n° 9002 du département des études économiques d'ensemble de l'INSEE*.

ERKEL-ROUSSE H., (1995), « Détection de la multicolinéarité dans un modèle linéaire ordinaire: quelques éléments pour un usage averti des indicateurs de BELSEY, KUH ET WELS », *Revue Statistiques Appliquées*, volume XLIII, numéro 4.

FOUCART F., (2006), « Colinéarité et régression linéaire », *Math. & Sciences. Humaines, ~ Mathematics and Social Sciences*, 44e année, n° 173, *2006*(1), p. 5-25.
http://www.ehess.fr/revue-msh/pdf/N173R963.pdf

FOUCART F., (2007) « Evaluation de la régression bornée » en cours de publication dans la *Revue des Nouvelles Technologies de l'Information*. Article consultable sur le site :
http://foucart.thierry.free.fr/colreglin/Regression_bornee.pdf

FREUND R.J., LITTELL R.C., (1991), *SAS System for regression*, 2nd edition, SAS-Editor.

GALTON F., (1886), «Regression towards mediocrity in hereditary stature», *Journal of the Anthropological Institute* 15 (1886), p246-263.
http://www.stat.ucla.edu/history/regression.gif

Greene W., (2005), *L'Econométrie*, Pearson Education », 5ème Edition
HOERL A.E., KENNARD R.W., (1970), « Ridge Regression: (1) biased estimation for non-orthogonal problems ; (2) applications to non-orthogonal problems», *Technometrics*, 12, pp. 55-67; pp. 68-82.

INDJEHAGOPIAN J.P., (1993), *Cours d'économétrie*, Polycopié ISUP.

LADIRAY D., (1990), « Autopsie d'un résultat : L'exemple des procédures Forecast, X11, Cluster », *Club SAS 1990*

LADIRAY D.,(1997 et suivantes), *Analyse exploratoire des données*, Cours polycopié de l'ENSAE.

LE GUEN M., (2001), « La boîte à moustaches de TUKEY, un outil pour initier à la statistique », *Statistiquement Vôtre*, n° 4, 14 pages.
http://matisse.univ-paris1.fr/leguen/leguen2001b.pdf

LE GUEN M., (2004), « L'analyse exploratoire des données et SAS/Insight. Visualisation dynamique des données », *Cahiers de la Maison des sciences économiques, Matisse, Série rouge*, n°2004.01, 13 pages,
ftp://mse.univ-paris1.fr/pub/mse/cahiers2004/R04001.pdf

MALINVAUD E., (1966), *Méthodes statistiques de l'économétrie*, Dunod

MOLES A., (1990), *Les sciences de l'imprécis*, Le Seuil

NETER J., WASSERMAN W., KUTNER M. H., (1990), *Applied Linear Statistical Models*, Irwin 3e édition

PALM R., IEMMA A.F., (1995), « Quelques alternatives à la régression classique dans le cas de la colinéarité », *Revue Statistiques Appliquées*, volume XLIII, numéro 2.

ROUSSEEUW P.J., LEROY A.M., (2003 -2e edition), *Robust regression and outlier detection*, Wiley.

SAPORTA G., (2006), *Probabilités, analyse des données et statistique*, Technip.

S.A.S , (1981), Technical Report A102, *SAS Regression Applications*

S.A.S , (1990), *Stat User's Guide version 6*

S.A.S , (1991), FREUND R.J., LITTELL R.C., *SAS System for regression*, (2ème édition)

SAUTORY O., (1995), *La Statistique Descriptive avec le Système SAS*, INSEE GUIDES numéros 1-2.

SAWA, T., (1978), «Information Criteria for Discriminating Among Alternative Regression Models», *Econometrica*, 46, 1273 - 1282.

SAVILLE J.D., WOOD G. R., (1990), *Statistical Methods: The Geometric Approach*, Springer-Verlag

SEN A., SRIVASTAVA M., (1990), *Regression Analysis, Theory, Methods, and Applications*, Springer-Verlag

STIGLER S. M., (1986), *The history of Statistics, The measurement of uncertainly before 1900,* The Belknap Press of Harvard University Press.

TENENHAUS M., (1994), *Méthodes statistiques en gestion*, Dunod-Entreprise.

TENENHAUS M., (1998), *La régression PLS : Théorie et pratique*, Editions Technip.

TENENHAUS M., GAUCHI J. P., MENARDO C., (1995) «Régression PLS et applications», R.S.A., volume XLIII, numéro 1.

TIBSHIRANI R., (1996), «Regression shrinkage and selection via the lasso», *J. Royal. Statist.* Soc B., Vol. 58, No. 1, pages 267-288,
http://www-stat.stanford.edu/~tibs/lasso/lasso.pdf

TOMASSONE R., ANDRAIN S., LESQUOY E., MILLIER C., (1992), *La Régression Nouveaux regards sur une ancienne méthode statistique*, INRA-Masson, 2e édition.

TOMASSONE R., DERVIN C., MASSON J. P., (1993), *Biométrie, modélisation de phénomènes biologiques*, Masson.

TUKEY J.W., (1977), *Exploratory Data Analysis*, Addison Wesley Publishing Company, Reading, Massachusetts.

WHITE H., (1980), *Econometrics*, volume 48, pages 817-838

WOOLDRIDGE J.M., (2000), *Introductory Econometrics: A Modern Approach*, South Western

WONNACOTT T.H., WONNACOTT R.J., (1991), *Statistique*, 4e édition, Economica.

YU CH. H0, animation sur le problème des multicolinéarités, http://www.creative-wisdom.com/multimedia/collinear.html, puis cliquer sur PC Version (17 megas) pour décharger la vidéo.